UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UNEMPLOYMENT DURATION DURING THE 2008 RECESSION: A STATISTICAL ANALYSIS OF THE CANADIAN EXPERIENCE

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

OF MASTER IN MATHEMATICS

BY

LENIN ARANGO CASTILLO

NOVEMBER 2014

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DURÉE EN CHÔMAGE PENDANT LA RÉCESSION DE 2008: UNE ANALYSE STATISTIQUE DE L'EXPÉRIENCE CANADIENNE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

LENIN ARANGO CASTILLO

NOVEMBRE 2014

[Cette page a été laissée intentionnellement blanche]

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep gratitude to my advisor, Professor Sorana FRODA, whose expertise, understanding, patience, and generosity added to my graduate experience. She spent a lot of time getting permission to access to the database used in this Master thesis. I appreciate her vast knowledge and skill in many areas as well as being open person to new areas.

For the academic, technical, and financial support, I thank the Department of Mathematics at the Université du Quebéc à Montréal, its professors and its staff.

I would also like to thank professor François WATIER and my supervisor for the time they spent writing reference letters on my behalf for my Ph. D. school application. I appreciate your support throughout this process.

I am grateful to professor Juli ATHERTON for her advice and encouragement. Moreover, I have greatly benefited from following her course in survival analysis.

Last, but by no means least, I thank my family and friends.

For any errors or inadequacies that may remain in this work, of course, the responsibility is entirely my own.

[Cette page a été laissée intentionnellement blanche]

CONTENTS

[Cette page a été laissée intentionnellement blanche]

LIST OF TABLES

LIST OF FIGURES

[Cette page a été laissée intentionnellement blanche]

ABSTRACT

This Master thesis analyses the impact of the financial crisis that started in 2008 on unemployment duration in Canada, using the unweighted Survey of Labour and Income Dynamics (SLID) database. We perform a survival analysis after conditioning on the number of unemployment periods. It turns out that the sets of characteristics considered to explain the unemployment duration differ when the period covering the financial crisis is compared with a period before the crisis. We find weak evidence that individuals who entered unemployment during the financial crisis stayed for a longer time in unemployment than those who entered unemployment in a period preceding the crisis.

Key words: financial crisis, unemployment duration, SLID, survival analysis, Kaplan-Meier estimator, Cox regression, Cox-Snell Residuals.

[Cette page a été laissée intentionnellement blanche]

RÉSUMÉ

Ce mémoire de maîtrise analyse l'impact de la crise financière qui a débuté en 2008 sur la durée au chômage au Canada, en se basant sur la base de données Enquête sur la dynamique du travail et du revenu (EDTR). On applique une analyse de survie en conditionnant selon le nombre de périodes de chômage. Il apparaît que l'ensemble de caractéristiques qui expliquent la durée au chômage est différent si on compare une période avant la crise avec une période qui couvre le début de la crise. Il y a aussi une faible indication que ceux qui sont allés au chômage pendant la crise sont restés plus longtemps au chômage que ceux qui sont entrés en chômage avant la crise.

Mots-clés : crise financière, durée au chômage, EDTR, analyse de survie, estimateur de Kaplan-Meier, régression de Cox, résidus de Cox-Snell.

[Cette page a été laissée intentionnellement blanche]

INTRODUCTION

How much time did individuals spend in unemployment during the financial crisis 2008? How different is the unemployment duration observed in the 2002-2007 time window from the one observed in the period 2005-2010? How does the duration in unemployment vary across individuals, region, age, gender, aboriginal status, visible minority status, immigration status and education levels? Answers to questions such as these are needed for several reasons. First, unemployment can be a very unjust and undemocratic punishment. Often it hits disadvantaged groups in society: the young, the unskilled, ethnic minorities or migrants. The long term unemployed not only lose their skills, they lose motivation, they fall ill: in crude economic terms human capital is being depreciated. Second, the welfare of the unemployed is more closely related to the time they spend without a job than to the fact of their being unemployed. In this sense, the usual official statistics i.e. the unemployment rate is a less useful statistic than the average duration in unemployment. Third, the length of unemployment spells plays a critical role in the economic theories of job search. The unemployment duration is an important variable which can explain the changes in labour markets and it is widely used in the job destruction and job creation models when analyzing the flows between employment, unemployment and out of the labour force.

In specific terms, unemployment duration refers to the amount of time that an individual remains unemployed. During a recession both the unemployment rate and the unemployment duration increase. This has consequences for household

spending and financial solvency which requires specific labour market policies.

The effects of the 2008 crisis are still being felt, six years on. The Gross Domestic Product (GDP) is still below its pre-crisis peak in many rich countries, especially in Europe, where the financial crisis has evolved into the Euro crisis. The effects of the crash are still rippling through the world economy. Canada's financial system has been relatively less affected by the global financial crisis than those of other industrialized countries such as the United States and Great Britain. However, while the Canadian financial system seems to be doing relatively better than those of other countries, Canada's economy is nonetheless feeling the global economic slowdown. The economic difficulties experienced by its largest trading partner -the US- are resulting in weaker Canadian exports and further problems for the manufacturing sector. Moreover, the strong Canadian energy and natural resources sector is likely to suffer as the world economic slowdown brings about lower demand and weaker prices for commodities (Bergevin, 2008).

Although Canada has been relatively sheltered from the worst of the crisis, the impact of the economic slowdown in the US has affected, and will continue to affect, Canadian economy. For example, Canada's labour market recovery is not yet complete after the financial crisis, although it continues to outpace that seen in many other countries (e.g. Spain, Greece, Ireland, the US, etc.) members of The Organisation for Economic Co-operation and Development (OECD). The unemployment rate, as defined by the International Labour Organization (ILO), was 7.1% in the first quarter of 2013, down from a peak of 8.5% in the third quarter of 2009. According to the 2013 OECD Employment Outlook, the recovery will continue in Canada bringing unemployment down to 6.7% by the end of 2014, still moderately above its pre-crisis level of 6.1% (OECD, 2013). In

2012, Statistics Canada reported that the unemployment rate of youths aged 15 to 24 was 14.3%, compared with a rate of 6.0% for workers aged 25 to 54 and workers aged 55 or older. The youth unemployment rate was 2.4 times that of workers aged 25 to 54, the biggest gap recorded since 1977.

As previously mentioned, the unemployment rate, while certainly being one of the most closely watched economic indicators, offers on its own a rather incomplete picture of the labour market. An unemployment rate of say, 5% may express two very different realities: i) one is a situation in which 5% of the labour force becomes unemployed each month and spends only a few weeks looking for a job, or ii) a case in which the same 5% of the labour force is unemployed for the entire year. In the first case the labour market is characterized by a great deal of fluctuations with spells of unemployment not having serious consequences, while in the latter we see a stagnant market with unemployment implying severe hardship. The implications of these two scenarios for the well being of the unemployed are very different. To accurately understand the situation requires a reliable indicator of the average duration of a spell of unemployment (Corak and Heisz, 1995).

For these reasons, in recent years, various survival analysis and duration techniques for modelling the length of unemployment spells and strike durations have gained popularity in the social sciences. This literature has drawn heavily on statistical methodology developed largely in industrial engineering and in the biomedical sciences. These methods have a natural application to many economic problems. For example, seminal papers such as *Econometric Methods for the Duration of Unemployment* (Lancaster and Nickell, 1980) or *Estimating the Probability of Leaving Unemployment* (Nickell, 1979) propose and apply hazard function methods for studying unemployment durations. Since these two

mentioned papers were published, survival analysis has become a more common technique used in economics.

In this thesis we apply the survival analysis methodology to some Canadian public datasets to answer the questions mentioned above; in particular we use non-parametric and semi-parametric methods. These analyses identify the statistical effect of explanatory variables, such as personal characteristics, on the exit rate out of unemployment. This in turn enables one to identify groups of individuals with higher expected durations.

Two public datasets from Statistics Canada are used in this thesis. They are: i) the Labour Force Survey (LFS), and ii) the Survey of Labour and Income Dynamics (SLID). The former is used as a general reference for some definitions and to understand the survey methodology. The latter contains the information on unemployment durations and the main covariates used in our analysis. Both databases are longitudinal and are organized by Panels. In this thesis, we use the Panel 4 and Panel 5 from the SLID. Panel 4 covers the period from January 1st, 2002 to December 31st, 2007, and Panel 5 covers the period from January 1st, 2005 to December 31st, 2010, i.e. it covers the peak year (2008) of the financial crisis.

Two different methods are used to organize the SLID dataset before applying the statistical methods mentioned above. In Method I we group the information per person and we compute the total duration unemployment per person during each Panel time window. Method II emulates Boudreau and Lawless (2006) and Hajducek and Lawless (2012) and we analyze the crude unemployment durations

as listed in the SLID dataset. In the following chapters we carefully explain each method. After treating the raw data provided by Statistics Canada we apply the following survival tools: Kaplan-Meier estimation, G-rho test, Cox model fitting (including deviance analysis and residual analysis). For confidentiality reasons (see Section 4.1.1) we use LOESS regression to smooth survival curves.

This thesis is organized as follows. In Chapter I we discuss the global crisis's effects and the variables that characterize the Canadian labour market. Chapter II discusses the unemployment definitions, introduces the SLID data, the data manipulation and descriptive analyses. In Chapter III we introduce the main statistical concepts used in our analysis. Chapter IV is devoted to a detailed presentation of our analysis and results. Finally, we comment our main results and we draw conclusions.

[Cette page a été laissée intentionnellement blanche]

CHAPTER I

THE GLOBAL CRISIS AND THE MAIN CHARACTERISTICS OF THE LABOUR MARKET

As mentioned above, in this Master Thesis we are studying the global crisis of 2008 through specific unemployment Canadian statistics. Namely we focus on data collected from 2002 to 2010.

It is well known that many economic data collected over time exhibit dramatic breaks in their behaviour, associated with events such as financial crises (Jeanne and Masson, 2000; Cerra, 2005; Hamilton, 2005) or abrupt changes in government policy (Hamilton, 1988; Sims and Zha, 2004, Davig, 2004). Of particular interest to economists is the apparent tendency of many economic variables to behave quite differently during economic downturns, when underutilization of factors of production (inputs to the production process) rather than their long-run tendency to grow governs economic dynamics (Hamilton, 1989, Chauvet and Hamilton, 2005). Time series of unemployment rates exhibit such behaviour.

In this section we discuss the effects of the Global financial crisis on Canada's economy as well as some institutional aspects of the Canadian labour market. Our purpose is to describe briefly the international context that motivates our

research, how the Canadian economy, in particular the labour market, has behaved under the recent Financial crisis, and some institutional aspects that give some advantage (or disadvantage) to Canada in its recovery.

1.1 The crisis of 2008: an overview

It is generally accepted that the acute phase of the current global financial and economic crisis started in September 2008, with the demise of Lehman Brothers. The causes (both short- and long-term) of the current global economic and financial crisis have been discussed in a number of contributions, including Aiginger (2009), Eichengreen and O'Rourke (October 2008), IMF (2008, 2009, 2010), Krugman (2008, 2009, 2010), Ormerond (2010), Solow (2009), UNCTAD (2008, 2009, 2010), UDESA (2010). Three main causes have been commonly accepted: deregulation in financial markets, world financial imbalance, and financial internationalization.

Before we go on to summarize the consequences of the financial crisis in Canada, we must understand the key features of economic data that define a business cycle. The term business cycle refers to economy-wide fluctuations in production, trade and economic activity in general over several months or years in an economy organized on free-enterprise principles. We refer to these patterns in fluctuations as a comovement.

One can see that business cycles are quite irregular, in that they are unpredictable; macroeconomic forecasters often find it difficult to predict the timing of a business cycle upturn or downturn. However, business cycles are quite regular in terms of comovements.

In Figure 1.1 we show the Gross Domestic Product (GNP) growth rate (in percentage, %) and the unemployment rate in Canada over the period 1995-2010[1] (OECD, 2013). There are peaks and troughs in the GDP growth rate; a series of positive deviations from the mean culminating in a peak represents a boom, whereas a series of negative deviations from the mean culminating in a trough represents a recession. The figure illustrates two important recent recessions, in the following years: 2001-02, and 2008-09.

While the GDP fluctuates in irregular patterns, macroeconomic variables exhibit strong regularities. In economic terminology, employment is a procyclical variable, that is, its deviations from the employment trend are positively correlated with the deviations from trend in GDP. Unemployment is countercyclical (negatively correlated). Figure 1.1 seems to confirm that unemployment rises when the GDP falls (OCDE, 2013).

A consequence of high unemployment is personal income decrease. In microeconomics terms, a decrease in personal wealth (measured by income *per-capita*) causes consumers to reduce their expenditures on normal and superior goods, which reduces cash flows for firms selling such goods and services and for suppliers of intermediate goods. With reduced demand for inputs of all sorts, employees are fired faster than they would be rehired elsewhere. Unemployment rose from 4.7% to 7.2% during 2007-2008 in the United States and from 6.1% to 8.3% in Canada during 2008-2009 (OECD, 2013). Therefore, it seems that the

[1]In Canadian dollars, in constant prices (national base year, previous year prices and OECD base year i.e. 2005). Expressed in millions.

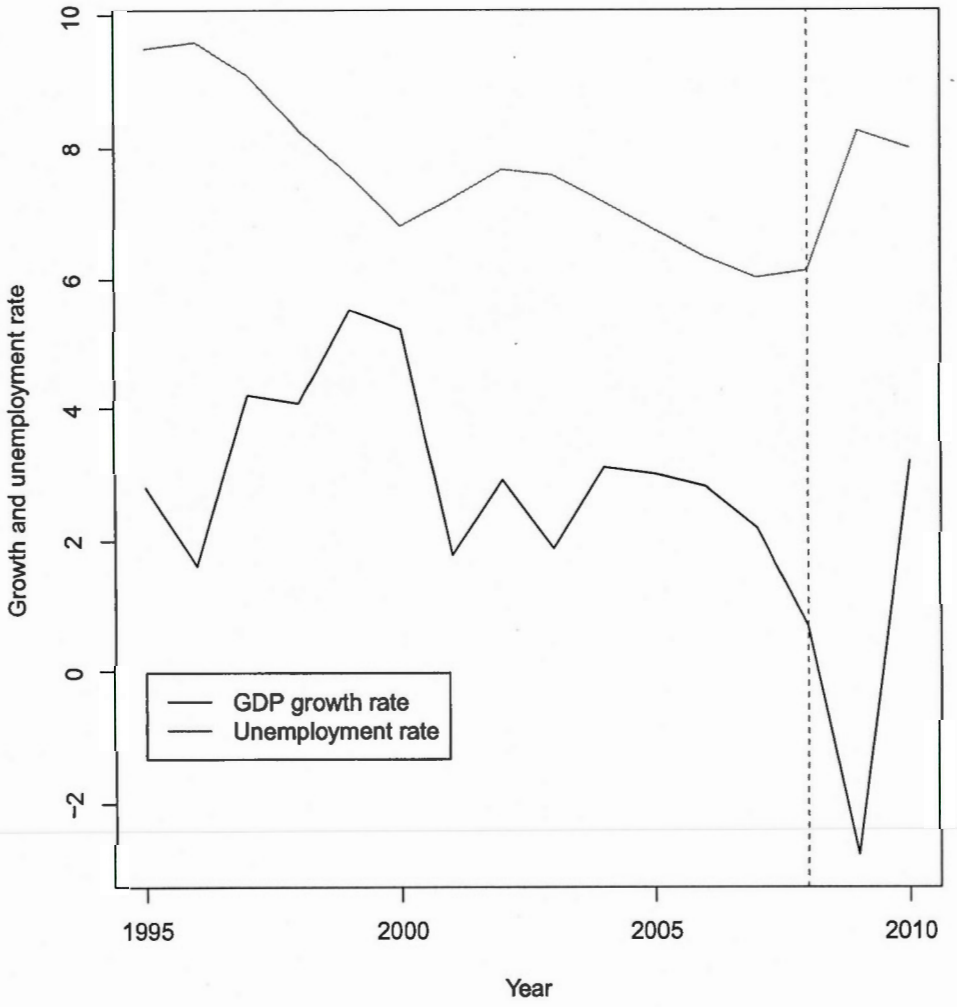


Figure 1.1: Canadian GDP real growth and unemployment rates (%)

consequences observed in the United States arrived in Canada one year later.

From Figure 1.1, it is important to note that the unemployment rate fell from 2002 to 2006 and it increased again from 2007 to 2010.

1.2 The effects of the Global Crisis on Canada and the Canadian response

As mentioned earlier, Canada went into recession later than the US and many other industrialized countries. However, Canada's recovery has seemed to evolve faster than in other OECD's members.

In Canada, one year after the beginning of the recession, the rate of job destruction (job losses) was as great as it was in the 1981-1982 recession and greater than in the 1990-1991 recession. As a result of the financial crisis, the business sector in Canada employed approximately 369 000 fewer individuals in 2009 compared to 2008 (Rollin, 2012). Part-time jobs were rising, and this is a sign of stress in the labour market. The official unemployment rate rose to 8.4% with more than one and a half million people looking for work.

Personal bankruptcies[2] climbed, as were credit defaults and house foreclosures. From 2007 to 2008, bankruptcies increased by 13.2% and from 2008-2009 they increased by 28.6%. According to the Office of the Superintendent of Bankruptcy Canada Annual Reports (2008, 2009), the province of Ontario was the most affected and reported 6 840 cases of personal bankruptcies in 2009.

The Canadian International trade activity slowed down abruptly. Between 2008 and 2009, exports fell by 120 thousands millions (CAD), which is equivalent to 24.0%. Even if the export recovery started in 2010, in 2012 Canada's export activity was still lower than at its level before the financial crisis. This fall in the

[2]The state of a consumer or a business that has made an assignment in bankruptcy or against whom a bankruptcy order has been made.

export activity was mainly driven by the lower demand from the US, reporting a fall around 100 thousand million (CAD) (Statistics Canada, CANSIM, table 228-0058, 2013).

The Canadian economy (measured by the GDP) shrank at a rate of 2.77% per year starting in 2009. This contraction was similar to the previous deep recession in 1981-1982 when the GDP fell by 2.86%. In Canada and the U.S. the fall in industrial output since 2008 recession is closely tracking the decline of industrial output following the 1930's.

However, by several economic measures, Canada was seen to be well-positioned to provide a more effective policy response entering this recession - on both the monetary and the financial front and fiscal front - than most other countries. The Bank of Canada intervened repeatedly during the recent financial crisis to provide large liquidity directly to the financial market participants in order to stabilize the financial system (Zorn *et al.*, 2009). Canada's comparative advantage derives from its strong banking system and the low government debt to GDP ratio. Although the latter gives to Canada fiscal room to implement an aggressive stimulus package, this capacity is meaningless unless it is used effectively.

The Canadian labour market characteristics and regulations play an important role in this recovery, essentially because of the factors given below.

a) The Public sector employment[3]. The split between the private sector and the public sector employment is an important aspect of labour market performance as the incentives, productivity, and performance of labour in the private sector are different from those in the public sector.

The compound annual growth rate (geometric mean) of total employees from 2000 to 2012 was 1.66%. For the public sector and the private sector the corresponding rates were 2.32% and 1.46% respectively. The public sector employees represent 26.00% and the private sector 84.00% of the total employees for the same period. However, the growth rates changed dramatically during the financial crisis: the total growth rate fell from 1.90% in 2008 to −2.35% in 2009. This fall was mainly driven by the fall in the growth rate of the private sector employees (1.18% in 2008 to −3.10% in 2009) while the public sector employees growth rate fell from 4.32% in 2008 to 0.08% in 2009.

A key difference between the two sectors is that governments are preoccupied with fulfilling social goals and objectives rather than pursuing economic or business objectives. In the public sector, political pressures often result in resources going to projects that are not in the best interest of society. In addition, government businesses tend to develop with less capital and thus are more labour-intensive than their counterparts in the private sector (Megginson and Netter, 2001).

Some researchers argue that a larger public sector leads to poorer outcomes in the labour market and, more broadly, to poorer economic performance (Amela *et al.*, 2012).

b) Minimum wages. Minimum wage legislation, one of Canada's oldest social policies, exists in every province and territory as part of employment standards legislation. The minimum wage is the lowest rate an employer can pay employees who are covered by the legislation.

Minimum-wage laws establish the lowest level of hourly pay that employers must legally pay their workers. Minimum wages have been shown to reduce employment opportunities for young and unskilled workers by restricting the ability of employers and employees to negotiate mutually beneficial contracts. In particular, minimum-wage legislation hinders low-skilled workers and new workforce entrants from negotiating for employment they might otherwise accept (Stigler, 1946; Palda, 2000). A large body of empirical research documents the adverse effects of high and increasing (over time) minimum wages, which include a reduction in employment.

According to the real minimum wage per hour database from OECD, from 2000 to 2012, Canada's average minimum wage is 6.71 USD and ranked 10th among 26 members[4]. Excluding the period 2000-2003, Canada exhibits a positive growth rate in minimum wages. Canada's position seems to be stable, i.e., from 2000 to 2012 Canada remained in the 10th position, except for 2010 when Canada dropped to the 11th position.

Data from Statistics Canada (Amela *et al.*, 2012) reveal that, in 2010, 58% of all minimum-wage workers in Canada were between the age of 15-24, of which 85.7% lived at home with their family. As an aside, let's note that higher minimum wages are associated with higher school-dropout rates, as the

[4]Real hourly and annual minimum wages are statutory minimum wages converted into a common hourly and annual pay period for the 26 countries for which they are available. The resulting estimates are deflated by national Consumer Price Indices (CPI). The data are then converted into a common currency unit using either USD current exchange rates or USD Purchasing Power Parities (PPPs) for private consumption expenditures.

increase in the minimum wage encourages teenage workers to leave school in search of employment.

c) Unionization. Another important structural element of labour markets is unionization. Unionization has been demonstrated to impede the flexibility of labour markets, a key factor necessary for good labour market performance.

A literature review on unionization and its economic effects (Aidt and Tzannatos, 2008) corroborates the finding of other studies. The authors concluded that union members and other workers covered by collective agreements receive, on average, wage premium over their non-unionized counterparts. Furthermore, the researchers noted that net profits, investment rates (physical capital), and spending on research and development tend to be lower in unionized than in non-unionized firms, even though unionized firms tend to adopt new technologies as fast as non-unionized firms.

Using the OECD database concerning trade union statistics, we found that the average rate of unionization in Canada, from 2000 to 2012, is around 29.9% using Administrative data and 27.8% using the Survey data[5]. Based on the average rate of unionization, Canada is in the 12th position among 33 OECD countries[6] for the same period.

---

[5]This analysis is based on the number of active trade union members and the number of wage and salary earners. Data on union membership are broken down by source of data (administrative or survey data). Membership corresponds to the number of wage and salary earners that are members of a trade union. Total number of wage and salary earners is taken from OECD Labour Force Statistics.

[6]Using the Administrative data as a source except for Australia and Mexico, for available

d) Other characteristics. All of the Canadian provinces have many other labour regulations including employment standard, occupation licensing, and worker's compensation.

   i) Employment Standard Acts. All Canadian provinces have their own standard acts. These acts have in summary two core features of provincial employment labour standard laws and codes: mandatory overtime and exemptions from minimum wages.

   ii) Occupation licensing. Regulation of occupation licensing can affect labour market performance by impeding the mobility of the worker. Occupation licensing governs the entry requirements need to hold job titles or to practice in such professional fields as medicine, law, accounting, engineering, electric technician, etc.

To summarize, Canada's economy has been touched by the current financial crisis. However, as we mentioned above, its recovery seems to evolve faster than in other OECD members. This recovery can be explained by the monetary and fiscal policies implemented by the Bank of Canada.

Despite the great performance during this difficult period, the labour market did not recover its pre-crisis level. In what follows, we are addressing this issue from a more refined point of view than the one based only on the unemployment rates.

---

data between 2000-2012.

Given the above analysis, we propose to study the Canadian unemployment portrait before and during the crisis in relation with the most commonly used variables for the Canadian labour market.

[Cette page a été laissée intentionnellement blanche]

CHAPTER II

THE DATABASES, DATA MANIPULATION, AND DESCRIPTIVE STATISTICS

The data for our empirical investigation are obtained from Statistics Canada. The main databases are "The Labour Force Survey" (LFS) and "The Survey of Labour and Income Dynamics" (SLID). The analysis spans the time between 2002 and 2010, namely, panels 4 and 5 of SLID. For each panel, unemployment duration and the variables gender, age, aboriginal background, immigration status, region, and education level were considered in our analysis.

This chapter is organized as follows: Section 2.1 describes the Labour Force Survey, Section 2.2 describes The Survey of Labour and Income Dynamics, in Section 2.3 we introduce the definition of the main variables in unemployment duration, and in Section 2.4 we describe the sample used in our analysis.

2.1 The Labour Force Survey (LFS)

Our analysis is based specifically on data from the SLID. However, because the SLID samples are selected from the monthly LFS it shares the latter's sample design. For this reason, it is important to discuss what the LFS is, and how it

is designed. In what follows we refer to the Guide to the Labour Force Survey (2013) available on the website of Statistics Canada.

The Labour Force Survey is a household survey carried out monthly by Statistics Canada. The interviews are taken during the so-called reference week, which is normally the week containing the 15th day of the calendar month. The labour force is composed by the civilian non-institutional population 15 years of age and over who, during the survey reference week, were employed or unemployed. The sampling unit is a household, which is, any person or group of persons living in a dwelling. A household may consist of any combination of: one person living alone, one or more families, a group of people who are not related but who share the same dwelling.

The objectives of the LFS have been to divide the working-age population into three mutually exclusive classifications - employed, unemployed, and not in the labour force - and to provide descriptive and explanatory data on members of each of these categories. The employed persons are those who, during the reference week did any work for pay or profit, or had a job and were absent from work. The unemployed persons are those who, during the reference week, were available for work and were either on temporary layoff, had looked for work in the past four weeks or were expecting to start a job within the next four weeks. Persons not in the labour force are those who, during the reference week, were unwilling or unable to offer or supply labour services under conditions existing in their labour markets, that is, they were neither employed, nor unemployed. Data from the survey provide information on major labour market trends such as shifts in employment across industrial sectors, number of hours worked, labour force participation (i.e. whether the person is or is not in the labour force), and

unemployment rates.

The LFS is the only source of monthly estimates of total employment, including the self-employment, full and part-time employment, and unemployment. It publishes monthly standard labour market indicators such as the unemployment rate, the employment rate and the participation rate. The LFS is a major source of information on the personal characteristics of the working-age population, including age, sex, marital status, educational attainment, and family characteristics.

Employment estimates include detailed breakdowns by demographic characteristics, industry and occupation, job tenure, and usual and actual hours worked. The survey incorporates questions permitting analyses of many topical issues, such as involuntary part-time employment, multiple job-holding, and absence from work. The LFS also provides monthly information on the wages and union status of the employees, as well as the number of employees at their workplace and the temporary or permanent nature of their job. Other demographic variables as belonging to a visible minority, immigration status, and aboriginal background are also included.

Unemployment estimates are produced by demographic group, duration of unemployment and activity before looking for work. Information on industry and occupation, and reason for leaving the last job is also available for persons currently unemployed or not in the labour market but with recent labour market involvement.

2.1.1 Survey methodology

The LFS is conducted nationwide, in both the provinces and the territories. **Its population coverage excludes**: persons living on reserves and other Aboriginal settlements in the provinces; full-time members of the Canadian Forces and the institutionalized population. These groups together represent an exclusion of approximately 2% of the population aged 15 and over.

National Labour Force Survey estimates are derived using the results of the LFS in the provinces. Territorial LFS results are not included in the national estimates and are published separately. Geographical location and confidential issues are the sources of this exclusion. The same is applicable for Indian Reserves.

For the purposes of sampling, the population in geographic areas (provinces and regions within provinces) is further partitioned into strata, in order to maximize the reliability of the estimates, while keeping collection costs at a minimum. Dwellings in strata are not selected directly. Small well-defined geographical areas called clusters are mapped across all parts of the 10 provinces. For example, in the 2006 Census, each cluster contains approximately 200 households. These clusters are used as the unit for stratification, as well as the unit for sample selection within stratum. A sample of clusters is selected in each stratum. All dwellings within selected clusters are listed and a sample of dwellings is chosen from each list.

The number of households sampled across the country has varied over the years as a result of varying levels of funding, and improvements in the survey design. Recently, the sample size has been approximately 56,000 households. The sample is allocated to provinces and strata within provinces in the way that best

meets the need for reliable estimates at various geographic levels. These include national, provincial, census metropolitan areas (large cities), economic regions, and employment insurance regions.

The LFS follows a rotating panel sample design; in which *households remain in the sample for six consecutive months.* The total sample consists of six representative sub-samples or panels, and each month a panel is replaced after completing its six month stay in the survey. Outgoing households are replaced by households in the same or a similar area. This results in a five-sixths month-to-month sample overlap, which makes the design efficient for estimating month-to-month changes. The rotation after six months prevents undue respondent burden for households that are selected for the survey.

Demographic information is obtained for all persons in a household for whom the selected dwelling is the usual place of residence. Labour force information is obtained for all civilian household members 15 years of age or older. Respondent burden is minimized for the elderly (age 70 and over) by carrying forward their responses for the initial interview to the subsequent five months in survey.

2.1.2 Data collection

Data collection for the LFS is carried out each month during the week following the LFS reference week. LFS interviews are conducted by telephone (Computer Assisted Telephone Interviews) by interviewers working out of a regional office site or by a personal visit from a field interviewer. The interviewer first obtains socio-demographic information for each household member and then obtains labour force information for all members aged 15 and over who are not members

of the regular armed forces. In subsequent monthly interviews the interviewer confirms the socio-demographic information collected in the first month and collects the labour force information for the current month.

2.2 The Survey of Labour and Income Dynamics (SLID)

The SLID is an important source of income data for Canadian families, households and individuals. It is a longitudinal survey and it provides an added dimension to traditional surveys on labour market activity and income: the changes experienced by individuals and families through time. Among the survey's key objectives is to understand Canadians' economic well-being.

The subjects for SLID are selected from the monthly LFS and thus share the latter's sample design. As mentioned above, the LFS sample is drawn from an area frame and is based on a stratified, multi-stage design. The total sample is composed of six independent samples, called rotation groups because each month one sixth of the sample (or one rotation group) is replaced. By definition, the SLID sample is composed of two panels. Each panel consists of two LFS rotation groups and includes roughly 17,000 households. A panel is surveyed for a period of six consecutive years. A new panel is introduced every three years, so two panels always overlap.

The SLID main difference with the LFS is that the former *interviews the same people from one year to the next for six years.* The survey's longitudinal dimension enables evaluation of concurrent and often related events.

SLID also provides information on a broad selection of human capital variables, labour force experiences and demographic characteristics such as education, family relationships and household composition.

Similar to the LFS, the SLID covers all individuals in Canada, excluding residents of Yukon, the Northwest Territories and Nunavut, residents of institutions and persons living on Indian Reserves or in military barracks.

2.2.1 Survey design

As mentioned above, the SLID sample is composed of two panels. Each panel consists of roughly 17,000 households and about 34,000 adults, and is surveyed for six consecutive years. A new panel is introduced every three years, so two panels always overlap.

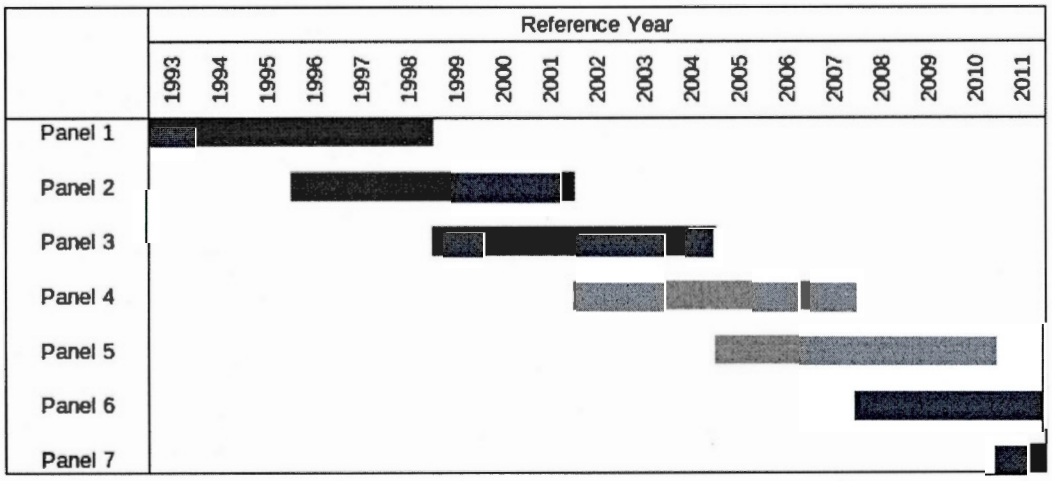


Figure 2.1: Overlapping design of SLID sample

Figure 2.1 shows the overlapping design of the SLID samples. For example, Panel 4 covers the period from January 1st, 2002 to December 31st, 2007, and Panel

5 covers the period from January 1st, 2005 to December 31st, 2010. These two panels overlap from January 1st, 2005 to December 31st, 2007. In our analysis, we analyze these two panels.

In a survey like SLID, the focus extends from static cross-sectional measures to a range of longitudinal events: transitions, durations, and repeat occurrences of people's financial and work situations. These yield a number of possible longitudinal researches themes.

Longitudinal respondents are the people belonging to the selected households when a new six-year panel of respondents is introduced. These respondents are interviewed once a year whether they stay, move away or split up. New participants, called cohabitants in SLID, are interviewed as long as they continue to live with a longitudinal respondent. That is because the family make-up and family income situation of the longitudinal respondents is of key interest. Interviewing cohabitants also improves the quality of cross-sectional estimates.

Children present in the original households are interviewed starting the year in which they reach 16 years. People aged 70 years and older are not asked labour-related questions.

Data collection follows the same process as the LFS mentioned in Section 2.1.2. Like the LFS, the SLID collects data on a wide range of topics. Some are inherently dynamic, involving transitions and spells, while others have important explanatory value. The main themes are listed below:

A) Labour

i) Activity

ii) Experience

iii) Jobless periods

iv) Job information

- Job characteristics
- Employer attributes
- Absences from work

B) Income

i) Income sources

ii) Employment insurance/Social assistance/Workers' compensation

C) Education

i) Educational activity

ii) Training

iii) Level of schooling

iv) Student loans

D) Personal characteristics

i) Demographics

ii) Ethnic-cultural

iii) Activity limitation

iv) Information on children

v) Geography

vi) Household and Family information

- Housing information

E) Sample control

i) Identifiers

ii) Weights

To comply with the strict confidentiality provisions of the Statistics Act, SLID longitudinal data are made available through special modes of dissemination. We had access to the SLID database through the Research Data Centre from Queen's University and through the Quebec Inter-University Center for Social Statistics from the branches in the Université de Montréal and the Université du Quebéc à Montréal.

2.3 Main definitions

The concepts of employment and unemployment are derived from the theory of the supply of labour as a production factor[1]. In this section we focus on general concepts and definitions of employment and unemployment used in this thesis.

According to the standard definition employed by Statistics Canada, the unemployed and the employed constitute the labour force. Detailed definitions for the variables used in this thesis will follow.

---

[1]The supply of labour is the total hours (adjusted for physical or mental intensity of effort) that workers wish to work at a given real wage rate. In economics, factors of production are the inputs to the production process.

Employment: Employed persons are those who, during the reference week:

a) did any work at all at a job or business, that is, paid work in the context of an employer-employee relationship, or self-employment. It also includes unpaid family work, which is defined as unpaid work contributing directly to the operation of a farm, business or professional practice owned and operated by a related member of the same household; or

b) had a job but were not at work due to factors such as own illness or disability, personal or family responsibilities, vacation, labour dispute or other reasons (excluding persons on layoff, between casual jobs, and those with a job to start at a future date).

Unemployment: Given the concept of unemployment as the unutilized supply of labour, the operational definition of unemployment is based primarily on the activity of job search and the availability to take a job. In addition to being conceptually appropriate, job search activities can, in a household survey, be objectively and consistently measured over time. The definition of unemployment is therefore the following.

Unemployed persons are those who, during the reference week:

a) were on temporary layoff during the reference week with an expectation of recall and were available for work, or

b) were without work, had looked for work in the past four weeks, and were available for work, or

c) had a new job to start within four weeks from the reference week, and were available for work.

Persons are regarded as available if they reported that: i) they could have worked in the reference week if a suitable job had been offered (or recalled if on temporary layoff); ii) if the reason they could not take a job was of a temporary nature such as: own illness or disability, personal or family responsibilities; iii) they already had a job to start in the near future; iv) they were on vacation (prior to 1997, those who were on vacation were not considered available). Full-time students currently attending school and looking for full-time work are not considered to be available for work during the reference week. They are assumed to be looking for a summer or co-op job or permanent job to start sometime in the future, and are therefore not part of the current labour supply.

Note that in the above definition there are two groups for which job search is not required: persons on temporary layoff and persons with a job to start at a definite date in the future. Persons on layoff are included among the unemployed on the grounds that their willingness to supply labour services is apparent in their expectation of returning to work. A similar argument is applied for persons who will be starting in a new job in four weeks or less.

The variables in our empirical analysis include age, gender, educational attainment, duration of unemployment, sex, immigration status, class of worker, and union status. In what follows we introduce the definitions of the terms and variables as used in the LFS and SLID as well as the variables selected from the SLID in our analysis.

Aboriginal identity: Persons who reported identifying with at least one Aboriginal group, for example, North American Indian, Métis or Inuit. This is

based on the individual's own perception of his/her Aboriginal identity, similar to the concept used with the Census. "Aboriginal identity" is not to be confused with "Aboriginal ancestry", another concept measured by the Census, but not by the LFS.

Age: Age is collected for every household member in the survey, and the information on labour market activity is collected for all persons aged 15 and over. Prior to 1966, information on labour market activity was collected for persons aged 14 and over. Beginning January 1997, date of birth is collected to ensure inclusion of respondents who turn 15 during their six month rotation in the survey.

Class of worker: There are two broad categories of workers: those who work for others (employees) and those who work for themselves (self-employed). The first group is subdivided into two classes: public sector employees and private sector employees.

a) The public sector includes employees in public administration at the federal, provincial, territorial, municipal, First Nations and other Aboriginal levels as well as in Crown corporations, liquor control boards and other government institutions such as schools (including universities), hospitals and public libraries.

b) The private sector comprises all other employees and self-employed owners of businesses (including unpaid family workers in those businesses), and self-employed persons without businesses.

Educational attainment: Highest level of schooling completed.

In this study we use the variable with the following categories: Less than high school graduation (1), High school diploma/degree (2), Non-university postsecondary certificate (3), and University degree or certificate (4).

Duration of unemployment: Number of continuous weeks during which a person has been on temporary layoff or without work and looking for work. Respondents are required to look for work at least once every four weeks, but they are not required to undertake job search activities each week in order to be counted as unemployed. A spell of unemployment is interrupted or completed by any period of work or withdrawal from the labour force.

In the SLID there exists the variable which indicates the total number of weeks in unemployment for each observation. We have also the start date and the end date for each period in unemployment. These three are used in our computation of unemployment duration.

Landed immigrant: Refers to people who are, or have been, landed immigrants in Canada. A landed immigrant is a person who has been granted the right to live in Canada permanently by immigration authorities. Canadian citizens by birth and non-permanent residents (persons from another country who live in Canada and have a work or study permit, or are claiming refugee status, as well as family members living here with them) are not landed immigrants.

Region. It indicates the region of residence for the household as of December 31 of the reference year. We analyse the following regions

i) Atlantic Canada (1). It is the region of Canada comprising the four provinces located on the Atlantic coast, excluding Quebec: the three Maritime Provinces - New Brunswick, Prince Edward Island, and Nova Scotia - and the east-most province of Newfoundland and Labrador.

ii) Quebec (2).

iii) Ontario (3).

iv) Prairies (4) The Prairie provinces or simply the Prairies comprise the provinces of Alberta, Saskatchewan, and Manitoba.

v) British Columbia (5).

vi) Category -other- comprises the Canadian territories: Yukon, Northwest territories, and Nunavut.

Union status: Beginning January 1997, employees are classified as to their union status: a) union member; b) not a member but covered by a union contract or collective agreement; or c) non-unionized.

Visible minority: It refers to whether a person belongs to a visible minority group as defined by the Employment Equity Act and, if so, the visible minority group to which the person belongs. The Employment Equity Act defines visible minorities as "persons", other than Aboriginal peoples, who are non-Caucasian in race or non-white in colour". The visible minority population consists mainly of the following groups: Chinese, South Asian, Black, Arab, West Asian, Filipino, Southeast Asian, Latin American, Japanese and Korean.

In what follows, we divide the covariates in two groups: i) the fixed covariates (sex, aboriginal background, immigration status, and visible minority), and ii) the dynamic covariates (education attainment and region). The levels considered for these factors in our thesis are summarized in Table 2.6 and Table 2.7.

We remark that, in the case of the fixed covariates, some other codes exist. These fixed covariates and the dynamic covariate related to education attainment share additional range and codes such as: 6 (Interim processing code), 7 (don't know), 8 (refusal), and 9 (not applicable). The dynamic covariate related to the region has additional codes such as 96 (Interim processing code), 97 (don't know), 98 (refusal), and 99 (not applicable).

The variable age is the only continuous variable. In this case, as for the fixed and dynamic covariates, there exist some additional codes such as: 997 (don't know), 998 (refusal), and 999 (not applicable).

For our independent variable, the number of weeks in unemployment, there also exist some additional codes named differently. These codes are: 9996 (Interim processing code), 9997 (don't know), 9998 (refusal), and 9999 (not applicable).

In this thesis, all additional codes to those mentioned in Table 2.6 and Table 2.7 are excluded from our analysis. However, it is important to mention them in order understand the following section.

2.4 The Sample, Covariates, and Descriptive analysis

For our analyses we were provided with the raw SLID sample data. Because our analyses are based on unweighted information, they are not comparable with the official statistics which are released by Statistics Canada. However, we believe that our results give an idea on how the financial crisis of 2008 impacted the Canada Labour Market.

2.4.1 Treatment of the raw data

As mentioned, we focus on the analysis of two panels to study the effect of the financial crisis on unemployment duration: Panel 4 and Panel 5. Our main interest is in considering individuals interviewed in these two panels. We study the total duration in unemployment for each individual which is a random variable, T.

From the Data Centre we received two unsorted datasets for each panel: i) jobless periods, and ii) job information. The first data set contains a list of unemployment periods and information on the people who declared to have been in unemployment during the reference month. It includes also variables such as age, sex, start and end date of the unemployment period, number of weeks in unemployment, aboriginal status, visible minority status, immigration status, level of education, and region. The second dataset contains information on all the persons in the SLID panels and includes such variables as class of worker and union status. Both datasets include the person's identifier (ID).

Thus, it is important to note that there is a difference between an observation and a subject (individual). For this reason, each person in unemployment may

appear more than once in the database, whenever the same person has been in unemployment more than once during the Panel time window.

In what follows we illustrate our data manipulations by considering three individuals (Table 2.1), who were followed in Panel 4 from time $t_1 = 2002$ to $t_6 = 2007$. The variables such as the person id (ID); the spell id (spellid); the main information for computing the duration in unemployment: start date (strdat7), end date (enddat7), and number of weeks in unemployment (nbwks). Further, one of the time independent fixed factors (Y_i), and the value of one of the dynamic factors ($Z_i(t_j)$) for person $i = 1, ..., n$ and year $j = 1, ..., 6$. We see that the same individual can be observed more than once if he experienced more unemployment periods. The start date in unemployment could be before January 1st 2002 (for person with ID=1, first observation), and there is some missing or incomplete information for some observations.

In Table 2.1 the person with ID=1 has been in unemployment for three times, the person whose ID=2 has also been three times in unemployment, and person with ID=3 has been unemployed only twice. For ID=1 the first observation (first unemployment period) has a starting date before January 1st, 2002. For person two, the first observation has a code 99979797 for the starting date meaning that we do not know when this unemployment spell started. As a consequence, we can not determine the number of weeks for this unemployment spell. For person with ID=3, we observe a code 6 on at least one of the fixed covariates, which means that Statistics Canada is still processing the information. Therefore, this person is completely dropped out of the analysis.

ID	spellid	strdat7	enddat7	nbwks	Y_i	$Z_i(t_1)$	$\cdots$	$Z_i(t_6)$
1	1	19951015	20020510	343	1	4	$\cdots$	4
1	2	20021215	20030901	37	1	4	$\cdots$	4
1	3	20050315	20060312	52	1	4	$\cdots$	4
2	1	99979797	20020101	9997	2	2	$\cdots$	3
2	2	20020502	20030131	39	3	2	$\cdots$	3
2	3	20051201	20071231	109	3	2	$\cdots$	3
3	1	20040112	20040901	33	6	1	$\cdots$	1
3	2	20051201	20061101	48	6	1	$\cdots$	1

Table 2.1: Data frame for duration times of individuals 1, 2, and 3

In what follows, we focus only on the labour database for each panel. Note that we use two different empirical methods commented on the Introduction and detailed in Chapter IV. For this reason, for each of our approaches we generated two different databases which are described in the following subsections.

Method I First, consider Panel 4. The Panel 4 includes around 27 958 observations (remember that each individual could have more than once unemployment period, and each unemployment period is one observation). For some observations, the information related to the jobless duration is not available or has not been processed. The codes for those cases were mentioned above and they were excluded in our analysis.

In the first step, observations with missing information for the starting date (8 195 observations) and with starting date before January 1st, 2002 (4 929 observations) were dropped out of our sample. In the case of Panel 4 (2002-2007),

this reduced the number of observations from 27 958 to 14 834. The main result of step 1 was that the maximum duration in unemployment for person $i = 1, .., n$ is 313 weeks.

In the second step, we added the label for the right censored observations (see Chapter III, Section 3.1.2 for details on censoring). To continue with the same example, let $C = 1$ denote the observations with complete unemployment duration and $C = 0$ denote the observations for which we do not know the exact unemployment duration ($C = 0$ indicate the right censored observations). Note that we kept only unemployment durations with start and end date (possibly censored) between January 1st 2002 to December 31st, 2007 (for the Panel 4 time window).

ID	spellid	strdat7	enddat7	nbwks	C	Y_i	$Z_i(t_1)$	$\cdots$	$Z_i(t_6)$
1	2	20021215	20030901	37	1	1	4	$\cdots$	4
1	3	20050315	20060312	52	1	1	4	$\cdots$	4
2	2	20020502	20030131	39	1	3	2	$\cdots$	3
2	3	20051201	20071231	109	0	3	2	$\cdots$	3

Table 2.2: Data frame for duration times of individuals 1 and 2 (Step two)

In addition, when there was missing information on at least one of the covariates, we dropped the observation (and consequently person in case of Method I). For example, the information available for person with ID=3 three is currently declared in process (Code 6 in Y_i) by Statistics Canada for at least one of the fixed covariates (Y_i). Hence, both observations for the person have missing information and both were dropped from our analysis. Table 2.2 shows the remaining data after step two. We remark that the person with ID=2 stays with

only two observations and its corresponding observation with spellid=3 is right censored.

In step three, we consider the **total duration in unemploymentt**. This was done as follows

i) We sorted the current database (after step two) using the person's identifier and by decreasing values of spellid;

ii) We counted the number of periods in unemployment inside the Panel time window in the current database. This sum is denoted by #UPer in the database in Table 2.3.

iii) We summed up the number of weeks in unemployment for each person during the Panel time window. We added this sum in the new variable named TDurU. Notice that TDurU denotes the total unemployment duration for each person remaining in our dataset.

After processing the raw information as mentioned above, the database looks as in Table 2.3. It contains the person identifier, the spell identifier, the independent variable (total duration in unemployment per person) and the set of fixed and dynamic covariates.

ID	spellid	strdat7	enddat7	nbwks	C	# UPer	TDurU	Y_i	$Z_i(t_j)$
1	3	20050315	20060312	52	1	2	89	1	4
1	2	20021215	20030901	37	1	2	89	1	4
2	3	20051201	20071231	109	0	2	148	3	2
2	2	20020502	20030131	39	1	2	148	3	2

Table 2.3: Data frame for duration times of individuals 1 and 2 (Step three). All values of $Z_i(t)$ not shown, $j = 1, .., 6$

Finally, in step four, we eliminated duplicates, i.e. we kept only one observation per person, that is, only one (cumulative) unemployment period. For example, in Table 2.3, the first line for person 1 and the first time for person 2 includes all the relevant information (the total number of weeks in unemployment per person, the right censored indicator, and the fixed and dynamic covariates).

Table 2.4 shows the final the database after step four. We kept only the information for person one and person two. For person one, we kept the following relevant information: the censoring information indicating that this total unemployment duration is complete for the Panel time window, the number of unemployment periods considered in our analysis, the total unemployment duration, and the fixed and dynamic covariates. For person two the same information is preserved, however, the total unemployment duration is right censored meaning that it is incompletely measured.

For Panel 4, these four operations reduced the sample to 7 544 unique persons with at least one jobless period between January 1st, 2002 and December 31st, 2007. All these persons have also reported all the information for the covariates

ID	spellid	C	# UPer	TDurU	Y_i	$Z_i(t_1)$	$\cdots$	$Z_i(t_6)$
1	3	1	2	89	1	4	$\cdots$	4
2	3	0	2	148	3	2	$\cdots$	3

Table 2.4: Data frame for duration times of individuals 1 and 2 (Step four)

considered in our analysis.

It is important to mention that the Tables 2.1 to 2.4 are given only as an illustration on the way we processed the information provided by Statistics Canada, in particular how we treated the raw data in order to obtain the total durations in unemployment per person necessary for our analysis. None of these tables reproduces some information in the actual database.

We repeated the same process for Panel 5. Panel 5 includes around 27 418 observations. As in the case of Panel 4, 7 425 observations do not contain information on the starting date of the jobless periods, and 5 554 had a starting date before January 1st, 2005. Therefore, in Panel 5 we reduce from 27 418 observations to 14 439 observations in step one.

After step three and four the sample for Panel 5, is reduced to 7 208 unique persons with at least one jobless period between January 1st, 2005 and December 31st, 2010.

Method II

In this case we carried out the analysis per observation instead of per person which is the main difference with Method I. This approach can be found in

Hajducek and Lawless (2012) and it will be described in Chapter IV. The dataset used in this case is treated in the same manner as the data set in Method I except there is a slight difference in the second step.

We began with the same step 1 in Method I. Table 2.5 shows the result from step two. Note that the person with ID=1 loses his first observation in step 1 because it did not belong to the chosen time frame. For the person with ID=2 we also dropped out the first observation because we do not know the starting date of the first unemployment spell. However, for this analysis, we renamed the remaining spellid's in the covariate spellid2 to account for lines of data that have been removed. In Table 2.5 we do not reproduce the strdat7 and enddat7 columns in order to allow for more space for the two new variables called spellid2 and Order, but the columns are present in the database. As in Hajducek and Lawless (2012), we use the factor Order to identify that an observation is a first unemployment duration in the time window span of each of the panels, that is, Order $= 0$ denotes a first unemployment period during the panel's time window and Order $= 1$ denotes an unemployment period of rank $k = 2, 3, ...$ in the time window covered by the panel.

ID	spellid	spellid2	Order	nbwks	C	Y_i	$Z_i(t_1)$	$\cdots$	$Z_i(t_6)$
1	2	1	0	37	1	1	4	$\cdots$	4
1	3	2	1	52	1	1	4	$\cdots$	4
2	2	1	0	39	1	3	2	$\cdots$	3
2	3	2	1	109	0	3	2	$\cdots$	3

Table 2.5: Data frame for duration times of individuals 1 and 2 (Method II). Start and end dates are not shown

These operations reduced the number of observations for Panel 4 to around 13 300 and for Panel 5 to 12 429 observations. Notice that for this analysis we merge the data in both Panels in a unique database resulting in a database of around 25 729 observations. To differentiate from Method I, we refer to the unemployment time duration only as **unemployment duration per observation**.

2.4.2 Covariates

The definitions of the covariates considered in this study are given in Section 2.3. As mentioned above, we divide them in two groups: fixed and dynamic covariates. Our purpose is to consider spells beginning before the financial crisis and not affected by it (Panel 4), and a panel most likely affected by the crisis (Panel 5) and study how the covariates affect the unemployment duration in both panels.

Factor	Level	Description
Sex	1	Male
	2	Female
Aboriginal background	1	Yes
	2	No
Immigrant	1	Yes
	2	No
Visible minority	1	Yes
	2	No

Table 2.6: Fixed covariates

The main variable under analysis is the duration in unemployment (for Method I and Method II) derived in section 2.4.1 and the right censored indicator variable. In order to analyze the unemployment durations, we included some relevant covariates that could affect it. Table 2.6 and Table 2.7 list the covariates

considered in our analyses; their selection followed some preliminary analysis and consideration of the variables that were reasonably accurate, complete and of economical interest, as explained in Chapter I. Note in Table 2.6 and Table 2.7 the covariates and their relevant levels (codes used by Statistics Canada) for our analysis by both Methods described in the Section 2.4.1.

Variable / Factor	Level	Description
Age	Continuous	Person's age
Region	1	Atlantic
	2	Quebec
	3	Ontario
	4	Prairies
	5	British Columbia
Level of Education	1	Less than high school graduation
	2	Graduated high school
	3	Non-university postsecondary certificate
	4	University degree or certificate

Table 2.7: Dynamic covariates

2.4.3 Descriptive Analysis

As explained at the beginning of this section, our samples include all the persons who have declared to be in unemployment at least once during the SLID interviews and whose first unemployment period started on January 1st, 2002 or after for Panel 4 and on January 1st, 2005 or after for Panel 5. In addition to the start date, we consider the persons whose set of covariates is complete. As we did in Section 2.4.1, in this section we give the main descriptive statistics of the data obtained by each treatment method.

Method I

For Panel 4 the average age is 38, while for Panel 5 it is 37 years approximately. The number of observations for the other covariates (factors), for the levels mentioned on Table 2.6 and Table 2.7, are given in Table 2.8 and 2.9. For the rest of this thesis, covariate and factor are used indistinctly.

Variable	Description	Panel 4	Panel 5
		n	n
Sex	Male	3 656	3 541
	Female	3 888	3 667
	Total	7 544	7 208
Aboriginal background	Yes	401	422
	No	7 143	6 786
	Total	7 544	7 208
Immigrant	Yes	703	730
	No	6 841	6 478
	Total	7 544	7 208
Visible minority	Yes	513	606
	No	7 031	6 602
	Total	7 544	7 208

Table 2.8: Number of subjects for the fixed factors

Table 2.8 summarizes the number of observations for the fixed factors commented above. One could think immediately that there is a strong correlation among the fixed covariates such as visible minority, immigration status, and aboriginal status. Note that the aboriginal persons are not considered as a visible minority

variable. Note also that members of communities considered in the visible minority variable are not necessary immigrants (see the definitions at the beginning of the chapter). Therefore, we considered it convenient to continue working with them.

Variable	Description	Panel 4 n	Panel 5 n
Region	Atlantic	1 566	1 598
	Quebec	1 345	1 237
	Ontario	2 126	1 968
	Prairies	1 787	1 761
	British Columbia	720	644
	Total	7 544	7 208
Level of Education	Less than high school graduation	1 346	1 240
	Graduated high school	2 554	2 478
	Non-university postsecondary certificate	2 414	2 256
	University degree or certificate	1 230	1 234
	Total	7 544	7 208

Table 2.9: Number of subjects for dynamic factors

Table 2.9 summarizes the number of observations for the dynamic factors. We remark that the region of Ontario includes more persons than other regions. The group of persons with high school diploma is larger in both panels.

Method II

As we are dealing with observations, it does not make sense to compute descriptive statistics for the covariate *age* or to divide the data for each Panel. Therefore, we

summarize the counting for the whole dataset (Panel 4 and Panel 5) as shown in Table 2.10 and Table 2.11.

Variable	Description	Database n_u
Sex	Male	13 068
	Female	12 661
	Total	25 729
Aboriginal background	Yes	1 502
	No	24 227
	Total	25 729
Immigrant	Yes	2 175
	No	23 554
	Total	25 729
Visible minority	Yes	1 753
	No	23 976
	Total	25 729
Order	0	14 752
	1	10 977
	Total	25 729

Table 2.10: Number of observations for the fixed factors (Method II) (n_U denotes the number of times out of work)

A new factor *Order* (Order=1 if the observation corresponds to a second or higher unemployment period, Order =0 otherwise) is added to the fixed covariates. Notice that more than 50% are of the observations are the first unemployment period.

In the dataset for Method II, the number of observations for males is greater than

for females. In Method II we control for the number of periods in unemployment using the factor Order. We observe in Table 2.11 that the number of observations, for the dynamic factors, is greater in Ontario and for the covariate Graduate high school.

Variable	Description	Database n_U
Region	Atlantic	6 000
	Quebec	4 455
	Ontario	6 975
	Prairies	6 123
	British Columbia	2 176
	Total	25 729
Level of Education	Less than high school graduation	4 125
	Graduated high school	9 260
	Non-university postsecondary certificate	8 126
	University degree or certificate	4 218
	Total	25 729

Table 2.11: Number of observations for dynamic factors (Method II) (n_U denotes the number of times out of work)

Other remarks

Finally, the information provided in the dataset for the theme Labour and job information is not considered. Remember that this dataset contains information on union status and the class of worker. At the beginning of our analysis, after running our initial models with them, we found evidence that something was wrong with these two covariates, but the reason was unclear. Later on we got feedback from Statistics Canada concerning the union status and the class of

worker (public or private) covariates revealing that this information is incomplete and incorrect for our purposes (i.e. these variables reveal the current status of the individuals, that is, unemployed people are by default non-unionized and do not work for public or private. Self employed people fall in the same category.) However, as we commented in Chapter I, these variables have an important value explaining why Canada recovers faster than other countries from the 2008 financial crisis. A combination of other variables provided in the SLID and LFS could help to carry on such an analysis but this goes beyond the scope of this thesis.

In the following chapters, we will describe the statistical methodology used to analyze this information.

[Cette page a été laissée intentionnellement blanche]

CHAPTER III

USEFUL METHODOLOGY: MAIN CONCEPTS IN SURVIVAL ANALYSIS

This chapter is based on the methodology and notation described by Collett (2003) and Klein and Moeschberger (2003). These books are essentially geared towards the medical and biology applications. Therefore, in our presentation we adapted the classic biostatistics terminology to the economic context under analysis.

Survival analysis deals with the study of data in the form of times (durations) from a well defined time origin until the occurrence of some particular event or end point. By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs. This event may be, for example: death, the appearance of a tumor, the development of some disease, cessation of smoking, leaving unemployment, and so forth.

In this chapter, we introduce the main concepts and techniques used in the analysis of our unemployment data set described in Chapter II. The chapter is organized as follows: in Section 3.1 we begin by defining the notation used in this Master Thesis and the two main functions of central interest in duration models.

In Section 3.2 we focus on the model estimation describing non-parametric methods (Kaplan-Meier estimation). In section 3.3 we discuss how a comparison of two groups is done using non-parametric estimation and hypothesis testing. In section 3.4 we present the most common semi-parametric method (Cox regression) commonly used in survival analysis. In the last Section, we discuss some problems with the direct application of such models to our data.

Before proceeding, let's note that, for simplicity, we illustrate the theory on the simplest case in our data base, subjects with one observed unemployment period. In Section 3.5 we are pointing to the specifics and difficulties in analyzing our data set.

3.1 Notation, special features and main functions in survival analysis

The actual survival time of an individual, t, can be regarded as the observed value of a random variable, T, which can take only non-negative values. In the medical literature, T is a random variable denoting the time until death and t is an observed survival time.

In this Master Thesis the event of interest is *leaving unemployment.* Therefore, T is a random variable denoting the time until a subject leaves unemployment. In what follows, we may alternate between the terms death or leaving the unemployment, and respectively still alive or still in unemployment.

The main feature of survival data that renders standard methods inappropriate is that survival times are frequently censored (incomplete in a specific way). The survival time of an event is said to be right censored when the end-point

of interest has not been observed for that individual. This may be because the data from a study are to be analyzed at a point in time where some individuals are still alive and thus time to death is unknown. The SLID data set described in Chapter II exhibits this phenomenon in each panel. For example, in Panel 5 where individuals are followed from January 1st, 2005 to December 31st, 2010, some individuals are still in unemployment at the end of the Panel's time coverage.

Alternatively, the survival status of an individual at the time of the analysis might not be known because that individual has been lost to follow-up. This also happens in the SLID data set. For example, after being selected to participate in the SLID and being reported to be unemployed, an individual may have moved to another country where communication between Statistics Canada and the individual was no longer possible. The only information available on the survival experience of that individual is the last date on which he or she was known to be in unemployment. This date may well be the last time that the individual reported to Statistics Canada to be in unemployment.

An individual who entered a study at time t_0 (calendar time) experiences the event (leaves unemployment) at time $t_0 + t$. However, t can be unknown, either because the individual is still in unemployment when the panel ends or because he or she has been lost to follow-up. If the individual was last known to be unemployed at time $t_0 + c$, the time c is called a censored survival time. This censoring occurs after the individual has entered into a study, that is, to the right of the last known survival time, and is therefore known as right censoring. The right-censored survival time is then shorter than the actual, but unknown, survival time. Other forms of censoring exist but they are not relevant to our analysis.

When assuming random right censoring, the usual model is as follows. For a specific individual under study, we assume that there is an unemployment duration T and a random or censoring time C, independent of T. The exact unemployment duration T of an individual will be known if, and only if, T is less than or equal to C. If T is greater than C, the individual is a survivor (still in unemployment), and his or her event time is censored at time C.

Four functions that characterize the distribution of T are the survival function (the probability of an individual surviving to time t); the hazard rate (or function), sometimes termed risk function (e.g., the chance an individual of age t experiences the event in the next instant in time); the probability density function (the unconditional probability of the event occurring at time t); and the mean residual life at time t (the mean time to the event of interest, given the event has not occurred at t). In what follows we give a mathematical description of all functions but the last one. There are other related functions (e.g. cumulative hazard, cumulative distribution function, etc).

3.1.1 Survivor function and hazard function

The actual survival time of an individual, t, can be regarded as the observed value of a continuous random variable $T \geq 0$. We regard T as a random variable with cumulative distribution function $F(t)$ given by

$$F(t) = \mathbb{P}(T \leq t), \tag{3.1}$$

and a probability density function $f(t) = dF(t)/dt$. The cumulative distribution function of T represents the probability that the survival time is less than some value t.

The survival function $S(t)$, is defined to be the probability that the survival time is greater than t. It is the complement of the cumulative distribution function of T, and thus is

$$S(t) = \mathbb{P}(T > t) = 1 - F(t). \tag{3.2}$$

The survival function can therefore be used to represent the probability that an individual stays in unemployment from some time origin to some time beyond t.

The hazard function is used to express the risk or hazard of death at some time t, and is obtained from the probability that an individual dies at time t, conditional on he or she having survived up to that time. Consider the probability that the random variable associated with an individual's survival time, T, lies between t and $t + \Delta t$, conditional on T being greater than or equal to t, written $\mathbb{P}(t < T \leq t + \Delta t | T > t)$. This conditional probability is then expressed as a probability per unit of time by dividing by the time interval, Δt, to give a rate. The hazard function, $h(t)$ is then the limiting value of this quantity, as Δt tends to zero and is given by

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{\mathbb{P}(t < T \leq t + \Delta t | T > t)}{\Delta t} \right\}. \tag{3.3}$$

From equation (3.3), $h(t)\Delta t$ is the approximate probability that an individual leaves unemployment in the interval $(t, t + \Delta t)$, conditional on that person having been in unemployment up to time t. For example, if the survival time is measured in days, $h(t)$ is the approximate probability that an individual, who is in unemployment on day t, leaves unemployment the following day. For this reason, the hazard function is often simply interpreted as the risk of death at time t (in the usual context).

A related quantity to the hazard function given in equation (3.3) is the cumulative hazard function $H(z)$, defined by

$$H(t) = \int_0^t h(s)ds = -\ln[S(t)]. \tag{3.4}$$

The following relations exist between $S(t)$, $f(t)$, $h(t)$

$$S(t) = \int_t^\infty f(s)ds = \exp\left[-\int_0^t h(u)du\right], \tag{3.5a}$$

$$f(t) = -\frac{d}{dt}S(t) = h(t)S(t), \tag{3.5b}$$

$$h(t) = -\frac{d}{dt}\ln[S(t)] = \frac{f(t)}{S(t)}. \tag{3.5c}$$

In particular, the relation between the hazard and the survival function given by equation (3.5a) will be repeatedly used in our application.

3.1.2 Censoring

There exist three forms of censoring, namely right censoring, left censoring, and interval censoring. However, for the characteristics of the SLID's data, we will only formally define right censoring which appears in our data set and which we account for in our analysis in Chapter IV.

Right censoring occurs when a subject leaves the study before the event of interest occurs, or the study ends before the event has occurred.

In our context, it is convenient to use the following notation. For a specific individual $i = 1, .., n$ under study, we assume that there is a unemployment time T_i and a random censoring time, C_i. The $T's$ are assumed to be independent and identically distributed with probability density function $f(t)$ and survival function $S(t)$. The exact time in unemployment T of an individual will be known

if, and only if, $T_i \leq C_i$. If $T_i > C_i$ the individual is a survivor, and his or her event time is censored at C_i. The data from this experiment can be conveniently represented by pairs of random variables (Y_i, δ_i), $i = 1, .., n$, where δ_i indicates whether the lifetime Y_i corresponds to an event ($\delta_i = 1$) or it is censored ($\delta_i = 0$), so $Y_i = \min(T_i, C_i)$. In our analysis, all $C_i \leq C^*$, $C^* = 313$ weeks.

In what follows, it is important to differentiate between observed times: $y_1, ..., y_n$, survival times (until the event, $t_1, ..., t_r$), and censored times: $c_1, ..., c_{n-r}$.

3.2 Non-parametric estimation

An initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times for individuals in a particular group. Survival data are conveniently summarized through estimates of the survival function and hazard function. These methods are said to be non-parametric because they do not require specific assumptions to be made about the underlying distribution of the survival times.

Suppose first that we have a single sample of survival times, where none of the observations are censored. When no observations are censored, the survival function $S(t)$, is the probability that an individual stays in unemployment for a time greater than t. This function can be estimated by the empirical survival function, given by

$$\hat{S}(t) = \frac{(\#\text{ of individuals with survival times} > t)}{(\#\text{ of individuals in the data set })}. \tag{3.6}$$

Equivalently, $\hat{S}(t) = 1 - \hat{F}(t)$, where $\hat{F}(t)$ is the empirical distribution function, that is, the ratio of the total number of individuals in unemployment at time t to

the total number of individuals in the study. We have that $\hat{S}(t) = 1$ for values of t before the first death time, and $\hat{S}(t) = 0$ after the final death time. Moreover, since the estimated survival function $\hat{S}(t)$ is constant between two adjacent death times, a plot of $\hat{S}(t)$ against t is a step function. The function decreases immediately after each observed survival time (when arranged in increasing order).

Among the non-parametric estimation methods which take censoring into account we find: i) the life-table estimator for estimating the probability of leaving unemployment of people from a given panel, ii) the Kaplan-Meier estimator for estimating the survivor function, and iii) Nelson-Aalen estimator for estimating the cumulative hazard. In this Master Thesis we use only the Kaplan-Meier estimator of the survival function for right censored data.

3.2.1 Kaplan-Meier (product-limit estimator)

This is the first step in our analysis of unemployment duration. To obtain a Kaplan-Meier estimate, a series of time intervals is constructed. However, each of these intervals is designed to be such that one death time is contained in the interval, and this death time is taken to occur at the start of the interval. Remember that by death time in our context we mean leaving unemployment time.

Let's suppose that there are n individuals with observed survival times y_1, y_2,...,y_n. Some of these observations may be right-censored, and there may also be more than one individual with the same observed survival time. We therefore suppose that in our data set there are r distinct values corresponding to times of leaving unemployment ("deaths") among the n individuals in our sample, where $r \leq n$. After arranging these death times in ascending order, the j-th one is denoted by $t_{(j)}$, $j = 1, ..., r$, and so the r ordered death times are

$t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. The individuals who are alive and not yet censored just before the time $t_{(j)}$, including those who are about to die at this time, form the risk set, and its size is n_j, while d_j is the number of individuals who die at $t_{(j)}$ $(j = 1, 2, ..., r)$. Since there are n_j individuals who are alive just before $t_{(j)}$ and there are d_j deaths at $t_{(j)}$, the probability that an individual dies at $t_{(j)}$ given that he was alive just before $t_{(j)}$ is estimated by d_j/n_j. The corresponding estimated probability of survival beyond $t_{(j)}$, conditional that an individual has survived up to $t_{(j)}$, is then $(n_j - d_j)/n_j$, $j = 1, .., r$.

It sometimes happens that there are censored survival times that occur at the same time as one or more deaths. In this case, the censored survival time is set to occur immediately after the death time when computing the values of n_j.

The probability of survival $\mathbb{P}(T > t_{(k)})$ can be written as the product $\mathbb{P}(T > t_{(k)}|T > t_{(k-1)})\mathbb{P}(T > t_{(k-1)}|T > t_{(k-2)}) \cdots \mathbb{P}(T > t_{(1)}|T > t_{(0)})\mathbb{P}(T > t_{(0)})$, with $\mathbb{P}(T > t_{(0)}) = 1$ and each factor can be estimated as described above. This leads to the Kaplan-Meier estimator of the survival function, which is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \qquad t_{(k)} \leq t < t_{(k+1)}, \qquad k = 1, 2, ..., r, \tag{3.7}$$

with $\hat{S}(t) = 1$ for $t < t_{(1)}$, and where $t_{(r+1)}$ is taken to be ∞. For $t \geq t_{(r)}$ we have two cases. If the largest observation is a censored survival time, t^*, say, then $\hat{S}(t)$ is undefined for $t \geq t^*$. On the other hand, if the largest observed survival time is an uncensored observation $t_{(r)}$, then $n_r = d_r$, and so $\hat{S}(t)$ is zero for $t \geq t_{(r)}$.

time	n.risk	n.event	survival	std.err
1	100	15	0.8500	0.0357
2	83	5	0.7988	0.0402
⋮	⋮	⋮	⋮	⋮
57	4	1	0.0584	0.0296
58	3	1	0.0389	0.0253

Table 3.1: The Kaplan-Meier estimate and its estimated starndard error: R output using the data set *hmohiv*

In fact, the limiting value of the Kaplan Meier estimate for fixed t is given by equation (3.2) when we assume that the number of intervals tends to infinity and their width tends to zero (see Collett, 2003, Chapter III). Table 3.1 shows, only for illustration, an output of a Kaplan-Meier estimator using the R package (see Klein and Moeschberger, 2003, Chapter IV, p.93 for a detailed example)[1]. The column survival denotes the $\hat{S}(t)$ estimate.

A plot of the Kaplan-Meier estimator of the survival function is a step-function, in which the estimated survival probabilities are constant between adjacent death times and decreases at each death time, but not at each censoring time. Figure 3.1 shows an example of the estimated survival function based on Table 3.1.

[1]The *hmohiv* data set is drawn from a study of HIV positive patients. The study examined whether there was a difference in survival times of HIV positive patients between those who had used intravenous drugs and those who had not. This data set has been taken from Introduction to SAS. UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/sas/notes2.

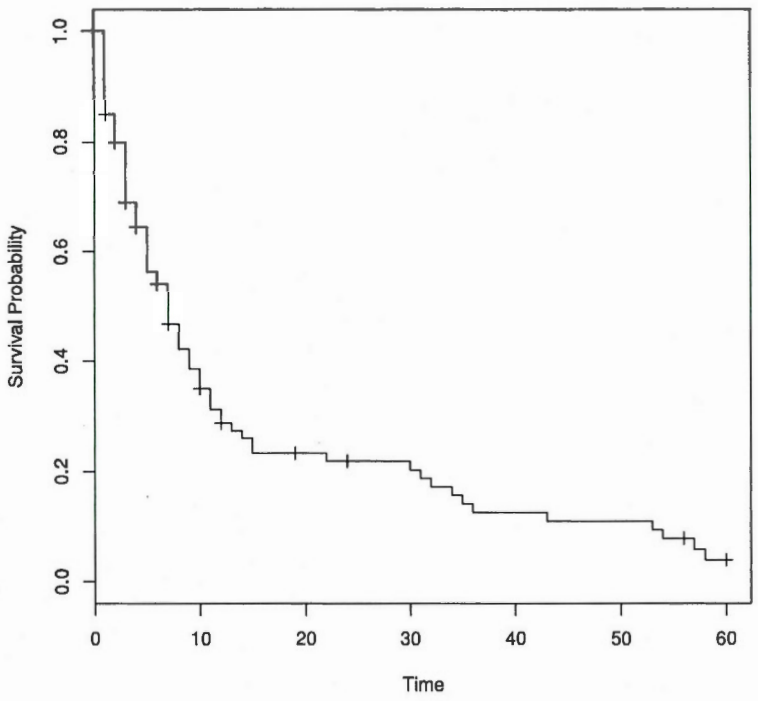


Figure 3.1: Estimated survival function: An example using the data set *hmohiv* and Table 3.1

3.3 Comparing the survival of two groups

In order to compare the unemployment duration between two different periods using Panel 4 and Panel 5, we need methods that allow us to compare two groups in terms of their survival data. There exists a variety of hypothesis testing procedures to carry on this comparison such as: the log-rank, Gehan, Tarone-Ware, Peto-Peto, modified Peto-Peto, or the Fleming-Harrington (see Collett, 2003, Chapter II and Klein and Moeschberger, 2003, Chapter VII).

In spite of the hypothesis testing procedures commonly used, there exists a simple way of comparing the survival times obtained from the two groups, in particular panels of individuals followed by the SLID. This is the comparison between the

plots of the two survival functions. The resulting survival plots can be quite informative regarding the probability to stay in unemployment and in which period this probability has been higher. We come back to this point in Section 3.4.1.

In Chapter IV, we use a combination of these plotting and testing procedures mentioned above. In the remainder of this section, we describe briefly the most commonly used hypothesis testing procedures.

3.3.1 Hypothesis testing procedures

Let τ be the largest time at which all of the groups have at least one subject at risk. This methodology is used in the SLID data to compare unemployment durations, for example, between Panel 4 versus Panel 5, between men versus women, or among different levels of education. In our case, $\tau = 313$ weeks.

Let M denote the number of groups under analysis. In general, $M = 2$ for our study, that is, when we compare individuals or observations from Panel 4 versus Panel 5. However M can be higher than two when we compare levels of education or regions.

In order to perform the comparison between M groups we can compare the hazard rates of $M \geq 2$ populations, that is, we test the following set of hypotheses

$$\begin{aligned} H_0: &\quad h_{(1)}(t) = h_{(2)}(t) = \cdots = h_{(M)}(t), \qquad \text{for all } t \leq \tau, \text{ versus} \\ H_A: &\quad \text{at least one of the } h_{(j)}(t)\text{'s is different for some } t \leq \tau. \end{aligned} \tag{3.8}$$

The inference on the hazard rates for all time points less than τ, which is,

typically, the smallest of the largest time on study in each of the M groups. The alternative hypothesis is a global one in that we wish to reject the null hypothesis, that is that, at least one of the populations differs from the others at some time.

Let $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ be the distinct death times in the pooled sample. At time $t_{(j)}$ we observe d_{jl} events in the lth sample out of n_{jl} individuals at risk, $j = 1, ..., r$, $l = 1, ..., M$. Let $d_j = \sum_{l=1}^{M} d_{jl}$ and $n_j = \sum_{l=1}^{M} n_{jl}$ be the number of deaths and the number at risk in the combined sample at time $t_{(j)}$, $j = 1, ..., r$.

The test of H_0 is based on weighted comparisons of the estimated hazard rate of the l-th population under the null and alternative hypotheses, based on the Nelson-Aalen estimator (Klein and Moeschberger, 2003). Namely, we compare the ratios d_{jl}/n_{jl}, $l = 1, ...M$ with d_j/n_j. If the null hypothesis is true, then, an estimator of the expected hazard rate in the l-th population under H_0 should be equal to the pooled sample estimator of the hazard rate d_j/n_j. Using data form the l-th sample, the estimator of the hazard rate is d_{jl}/n_{jl}. To make comparisons, let $W_l(t)$ be a positive weight function with the property that $W_l(t_j)$ is zero whenever n_{jl} is zero. The tests of H_0 are based on statistics of type

$$Z_l(\tau) = \sum_{j=1}^{r} W_l(t_j) \left[\frac{d_{jl}}{n_{jl}} - \frac{d_j}{n_j} \right], \qquad l = 1, ..., M. \tag{3.9}$$

If all $Z_l(\tau)$ are close to zero, then, there is little evidence that the null hypothesis is false. If one of the $Z_l(\tau)$'s is far from zero, then there is evidence that at least one population has a hazard rate differing from that expected under the null hypothesis.

By considering different weights we obtain different estimators. Harrington and

Fleming (1982) proposed a systematic estimator of weights

$$W_{p,q}(t_j) = \hat{S}(t_{j-1})^p \left[1 - \hat{S}(t_{j-1})\right]^q, \qquad p \geq 0, \;\; q \geq 0, \tag{3.10}$$

where $\hat{S}(t)$ is the Product-Limit-Estimator (i.e. equation (3.7)) based on the combined sample. It is a general class of tests that includes, as special cases, the log-rank test (see Collett, 2003, Chapter II) and a version of the Mann-Whitney-Wilcoxon test (see Klein and Moeschberger, 2003, Chapter VII). In equation (3.10), the survival function at the previous death time is used as a weight to ensure that these weights are known just prior to the time at which the comparison is to be made. Note that $S(t_{(0)}) = 1$ and we define $0^0 = 1$ for these weights. Special cases are:

i) $p = q = 0$ give the log-rank test;

ii) $p = 1$, $q = 0$ give the Mann-Whitney-Wilcoxon test;

iii) $q = 0$ and $p > 0$, these weight give the most weight to early departures between the hazard rates in the M populations;

iv) $p = 0$ and $q > 0$, these weights give most weight to departures which occur late in time.

By an appropriate choice of p and q, one can construct a test which has most power against alternatives which specify that the M hazard rates differ over any desired region.

In our statistical applications, we implemented a so-called $G - rho$ test which corresponds to $W_{p,q}(t_j)$ with $p = 1$ and $q = 0$, i.e. we applied an adapted version of the Mann-Whitney-Wilcoxon test.

3.4 Semi-parametric estimation

In this Master Thesis, as in most medical studies, we have supplementary information recorded on each individual surveyed by the SLID. As described in Chapter II, some socio-economic variables may have an impact on the time that the individual stays in unemployment. In order to explore how the survival experience of individuals is related to explanatory variables, an approach based on statistical modelling can be used.

In the analysis of survival data, interest centres on the risk or hazard of death at any time after the time origin of the study. As a consequence, the hazard function is modelled directly in survival analysis. The resulting models are somewhat different from linear models but similar to generalized linear models where the dependence of some function of the mean on certain explanatory variables is modelled. Actually, many of the principles and procedures used in generalized linear modelling carry over to the modelling of survival data.

There are two broad reasons for modelling survival data: i) to determine which combination of explanatory variables affect the hazard function, and ii) to obtain an estimate of the hazard function itself for an individual. The basic model for survival data to be considered in this thesis is the proportional hazards model. This model was proposed by Cox (1972) and has come to be known as the Cox regression or proportional hazards model. Although the model is based on the assumption of proportional hazards, no particular form of probability distribution is assumed for the survival times. The model is therefore referred to as a semi-parametric model. We now go on to develop the model for the comparison of the hazard functions for two individuals in Panel 4 and Panel 5.

3.4.1 The Cox model: main idea

Suppose that the hazard of death at a particular time depends on the values $x_1, ..., x_p$ of p explanatory variables, $X_1, ..., X_p$. The values of these variables will be assumed to have been recorded at the time origin of the study.

The set of values of the explanatory variables in the proportional hazards model will be represented by the vector $\mathbf{x}$, so that $\mathbf{x} = (x_1, ..., x_p)^\top$. Let $h_0(t)$ be a baseline hazard function for an individual, e.g., for whom the values of all the explanatory variables that make up the vector $\mathbf{x}$ are zero. The hazard function for individual $i = 1, ..., n$ can then be written as

$$h_i(t) = \psi(\mathbf{x}_i)h_0(t) \tag{3.11}$$

where $\psi(\mathbf{x}_i)$ is a function of the values of the vector of explanatory variables for the ith individual. The function $\psi(\cdot)$ can be interpreted as the hazard at time t for an individual whose vector of explanatory variables is $\mathbf{x}_i$, relative to the hazard for an individual for whom $\mathbf{x} = 0$ since $\psi(\mathbf{x}_i) = h_i(t)/h_0(t)$.

Since the relative hazard, $\psi(\mathbf{x}_i)$, cannot be negative, a natural idea is to write it as $\exp(\eta_i)$, where η_i is a linear combination of the p explanatory variables in $\mathbf{x}_i$. Therefore, let

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} = \sum_{j=1}^{p} \beta_j x_{ji} = \boldsymbol{\beta}^\top \mathbf{x}_i. \tag{3.12}$$

The general proportional hazards model then becomes

$$h_i(t) = \exp\left\{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right\} h_0(t). \tag{3.13}$$

Since the model can be re-written in the form

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}, \tag{3.14}$$

the proportional hazards model may also be regarded as a linear model for the logarithm of the hazard ratio. Notice that there is no constant term in the linear component of the proportional hazards model. If a constant term β_0, say, were included, the baseline hazard function could simply be rescaled by dividing $h_0(t)$ by $\exp(\beta_0)$, and the constant term would cancel out. Moreover, we have made no assumptions concerning the actual form of the baseline hazard function $h_0(t)$. Indeed, we will see later that the β-coefficients in this proportional hazards model can be estimated without making any such assumptions.

In view of the relation between $h(t)$ and $S(t)$ (see equation (3.5a)) one can see the following: assume that for some $i = 1, ..., n$ and all $t > 0$

$$h_i(t) = \gamma_i h_0(t), \qquad \text{with } \gamma_i < 1. \tag{3.15}$$

Then $S_i(t) = S_0^{\gamma_i}(t)$ and thus

$$S_i(t) > S_0(t), \qquad 0 < t < \infty, \tag{3.16}$$

and the survival functions $S_i(t)$ and $S_0(t)$ do not intersect. The inequality in Equation (3.16) will be repeatedly used in interpreting the results of our data analysis. In particular, under the Cox model, let γ_i take only two values ($\log \gamma_i = 0$ or $\log \gamma_i < 0$ and fixed) as in the case where we compare two groups. Then, the survival curve of one group lies always above the other one and they do not intersect.

There are two types of variables on which a hazard function may depend, namely variates and factors. A variate is a variable that takes numerical values that are often on a continuous scale of measurement, such as age in our study. A factor has no numerical meaning, but can be coded by a limited set of values, which are

known as the levels of the factor. Examples of factors are: sex, immigrant status, visible minority, etc..

Note that the Cox model can be fitted using only variates, only factors, or a combination of both. These models can also consider only main effects or interaction effects between a variate and a factor, or between two factors. Note that when a model with interaction is fitted, we do not omit the main factor/variate effect. The interactions can be of different orders, however, for this Master Thesis, we only consider interactions of order two.

3.4.2 Fitting the proportional hazard model

Suppose that data are available for n individuals, among whom there are r distinct death times and $(n - r)$ right-censored survival times. We will for the moment assume that only one individual dies at each death time, so that there are no ties in the data. The r ordered death times will be denoted by $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, so that $t_{(j)}$ is the jth ordered death time. The set of individuals who are at risk at time $t_{(j)}$ will be denoted by $R(t_{(j)})$, so that $R(t_{(j)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The quantity $R(t_{(j)})$ is called the risk set.

Cox (1972) showed that the relevant likelihood function for the proportional hazard model in equation (3.9) is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^\top \mathbf{x}_{(l)})} \tag{3.17}$$

in which $\mathbf{x}_{(j)}$ is the vector of covariates for the individual who dies at the jth ordered death time, $t_{(j)}$, $i = 1, ..., r$. The summation in the denominator of this

likelihood function is the sum of the values of $\exp(\boldsymbol{\beta}^{\top}\mathbf{x})$ over all individuals who are at risk at time $t_{(j)}$. Notice that the product is taken over the individuals for whom death times have been recorded. Individuals for whom the survival times are censored do not contribute to the numerator of the log-likelihood function but they do enter into the summation over the risk sets at death times that occur before the censored time. Moreover, the likelihood function depends only on the ranking of the death times, since this determines the risk set at each death time. Consequently, the inference about the effect of the explanatory variables on the hazard function depends only on the rank order of the survival times.

Now suppose that the data consist of n observed survival times, denoted by $y_1, y_2, ..., y_n$, and that δ_i is an event indicator, which is zero if the ith survival time, y_i, $i = 1, 2, ..., n$, is right-censored, and unity otherwise. The likelihood function in equation (3.11) can then be expressed in the form

$$\prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}^{\top}\mathbf{x}_{(i)})}{\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{\beta}^{\top}\mathbf{x}_{(l)})} \right\}^{\delta_i} \tag{3.18}$$

where $R(t_i)$ is the risk set at time y_i. The corresponding log-likelihood function is given by

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \exp(\boldsymbol{\beta}^{\top}\mathbf{x}_{(j)}) - \log \sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^{\top}\mathbf{x}_{(l)}) \right\}. \tag{3.19}$$

Numerical methods are used to compute the maximum likelihood estimates of the vector of β-parameters. In the case of multiple deaths at a given time one can perform a similar analysis by considering a more general formula for the likelihood function (see Chapter III in Collett (2003)).

3.4.3 Residual analysis

Many model-checking procedures are based on quantities known as residuals. These are values that can be calculated for each individual in the study, and have the feature that their behaviour is known, at least approximately, when the fitted model is satisfactory. A number of residuals have been proposed for use in connection with the Cox regression model; however, in this thesis we rely on one of the most commonly used in the analysis of survival data: the Cox-Snell residuals.

Throughout this section, we will suppose that the survival times of n individuals are available, where r of these are death times and the remaining $(n-r)$ are right-censored. We further suppose that a Cox regression model has been fitted to the survival times and that the linear component of the model contains p explanatory variables, X_1, X_2,....,X_p. The fitted hazard function for the ith individual, $i = 1, 2, ..., n$, is therefore

$$\hat{h}_i(t) = \exp\left(\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i\right)\hat{h}_0(t), \tag{3.20}$$

where $\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i$ is the value of the fitted component, or linear predictor, of the model for that individual and $\hat{h}_0(t)$ is the estimated baseline hazard function.

The Cox-Snell residual for the ith individual, $i = 1, 2, ..., n$, is given by

$$r_{C_i} = \exp\left(\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}_i\right)\hat{H}_0(t_i), \tag{3.21}$$

where $\hat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i (see equation (3.4)). Considering the relation between the survivor and the hazard function, the Cox-Snell residual, r_{C_i} is the value of $\hat{H}_i(t_i) = -\log \hat{S}_i(t_i)$, where

$\hat{H}_i(t_i)$ and $\hat{S}_i(t_i)$ are the respective estimated values of the cumulative hazard and survival function of the ith individual at t_i (see in Collett (2003), Chapter III).

The idea behind this type of residual analysis is the following. Let $S(t)$ be the true survival function of a random variable T. Then the random variable $S(T)$ has a uniform $U(0,1)$ distribution, and $Y = -\log S(T)$ has an exponential distribution with unit mean, irrespective of the form of $S(t)$.

The next and crucial step in the argument is as follows. If the model fitted to the observed data is satisfactory, then a model-based estimate of the survival function for the ith individual at t_i, $i = 1, ..., n$, where t_i is the survival time of that individual, will be close to the corresponding true value $S_i(t_i)$. This suggests that if the correct model has been fitted, the values $\hat{S}_i(t_i)$ will have properties similar to those of $S_i(t_i)$. Then, the negative logarithms of the estimated survival functions, $-\log \hat{S}_i(t_i)$, $i = 1, 2, ..., n$, should behave approximately as n observations from a unit exponential distribution of survival function $\exp(-t)$.. These estimates are the Cox-residuals and are implemented in Chapter IV to the SLID data set.

3.5 Remarks on our data

We described generally the main points of our methodology in the previous sections. However, applying this methodology to the SLID data set it is not as simple as we described above. Some problems emerge when we consider our data for analysis.

i) The methods described above assume that each individual is observed for a unique continuous period starting at some point after the beginning of some

study (clinical, economic, etc.). Commonly, individuals under observation experience the event only once and we register only one T survivor time per individual. The random variable T is assumed to be continuous of cumulative distribution function $F(t)$. In our case, many individuals experience more than one period in unemployment, that is, we have a multi-event process where during the panel's time window individuals could leave more than once from unemployment.

ii) As our objective is to compute the effect of the crisis in Canada during the 2008 financial crisis, we expect to assess this effect on the total duration in unemployment experienced by each individual in our data set. Therefore, we sum up all the unemployment periods of each individual in our data set and our random variable of interest is a sum of a random number of terms. Applying directly the methodology mentioned above could be problematic because, among others, two individuals who experience a different number of unemployment periods cannot be assumed to have a total time T following the same distribution. We need to recall that in the SLID data set each observation corresponds to one unemployment duration, and thus it is a "survival" time as defined in the standard survival literature. In the literature, in order to control for the fact that some individuals experience more than one unemployment period, an Order factor was usually added to the survival analysis (Order=1 if the observation corresponds to a second or higher unemployment period). Still, this approach does not take into account how unemployment is experienced at the individual level and the inherent dependence between such observations. We will see in Chapter IV that the two approaches give very different results.

iii) On the other hand, in each Panel, one can consider a mixture distribution

for the total time T, such that, for individual j, $T_j = \sum_{i=1}^{K_j} T_{i,j}$ where K_j can take one of the values 1, 2, 3, 4 (as we considered up to 4 unemployment periods). Thus, we could consider

$$\mathbb{P}(T_j > t) = \sum_{k=1}^{4} \mathbb{P}\left(\sum_{i=1}^{K_j} T_{i,j} > t | K_j = k \right) \mathbb{P}(K_j = k). \tag{3.22}$$

Still, in comparing two panels via a Cox model e.g., one should suppose that the two mixture distributions are of the same type, in particular $\mathbb{P}(K_j = k)$, $k = 1, 2, 3, 4$ are the same.

[Cette page a été laissée intentionnellement blanche]

CHAPTER IV

DATA ANALYSIS

In this chapter, we develop techniques for drawing inferences about the distribution of a particular time to event, based on our sample which has right censored data.

As mentioned above, our main variable under analysis is the **total duration in unemployment**. Since some subjects have more than one unemployment period, typically this time is a sum of random variables on a number of terms that depends on each subject. As mentioned previously, standard survival analysis methods do not necessarily apply, given that there are intermittent observation times, among other. So, in order to apply some standard techniques on duration times we decided to adopt two approaches: i) we propose an analysis for the *total duration in unemployment per individual* after dividing the sample into more homogeneous subgroups in the database, as treated by Method I, given in Chapter II, section 2.4.1, and ii) we carry a second analysis based on the *duration in unemployment per observation*, where we use the original database, as described in Chapter II, section 2.4.1. This last type of approach can be found in Boudreau and Lawless (2006) and Hajducek and Lawless (2012).

More precisely, we apply the following methodology:

i) in the case of Method I we use the dataset as derived in Chapter II after step four. Thus, the duration in unemployment $T_j, j = 1, 2, \ldots, n$ indicates the total unemployment duration for each individual, $j = 1, 2, \ldots, n$. We treat this total time as right censored if at the end of the time window covered by each panel, the person is still in unemployment. Note that T_j is a sum of a random number of terms, $T_j = \sum_{i=1}^{k_j} T_{i,j}$, where $T_{i,j}$ is the $i - th$ duration in unemployment for individual j (with $i \leq k_j$) and k_j is the number of unemployment periods of subject $j = 1, 2, \ldots, n$. So, as a first analysis, we divide the subjects according to the number of unemployment periods, $k = 1, 2, 3, 4$, and we compare the cohorts corresponding to each value of k in Panel 4 and Panel 5;

ii) in the case of Method II we replicate, at our best, the methodology proposed by Hajducek and Lawless (2012). The dataset by observation as described in Chapter II, Method II, is used in our computations with a slight difference from the afore-mentioned authors. The difference consists in the fact that we eliminate observations where the starting date in unemployment precedes the starting date of the respective Panel. In this case the duration in unemployment $D_\ell, \ell = 1, 2, \ldots$ indicates a duration corresponding to an observation in the data base, and not to a specific individual. Thus, we are ignoring that the same person could correspond to more than one unemployment period in the data base. Emulating Hajducek and Lawless (2012) the statistical analysis comprises an additional factor (named Order), at two levels, which controls for the number of unemployment periods in a certain way. Its levels are: Order = 0 indicates that the respective time is a first unemployment period, while Order = 1 indicates that the time is a second, third, etc, unemployment period.

In addition to our main analysis methodologies presented in points i) and ii), in section 4.3 we analyse both panels independently, that is, we focus on finding the main explanatory variables of unemployment duration for persons belonging to Panel 4 and Panel 5, when controlling for the number $k = 1, 2, 3$ of unemployment periods.

Factor	Baseline Category	Codes
Sex	Man	Sex
Aboriginal Background	Yes	Abor. Backg.
Immigrant	Yes	Immigrant
Visible Minority	Yes	Vis. Min.
Order	First unemployment period	Order
Panel	Panel 4	Panel5
Education	No high school diploma	Educ2
		Educ3
		Educ4
Region	Atlantic	QC
		ON
		AB, SK, MB
		BC

Table 4.1: Factors, baseline categories and codes for categories other than the baseline

In the following sections we analyze the unemployment duration for both approaches using appropriate survival analysis techniques. One aim of our analysis is to determine whether the Cox model is applicable, and to take into account the

explanatory variables given in Chapter II. Wherever the Cox model is applicable, we give the main findings concerning the effects of covariates in the unemployment duration per person or per observation. We carry also an analysis for each Panel in order to determine the covariates that have affected the unemployment duration for the time window covered by each panel. Following the estimation, we perform also a residual analysis in order to assess the validity of the proposed models.

Before proceeding, in Table 4.1 we describe how the explanatory factors were coded. We considered a continuous variable (age), five factors at two levels (sex, aboriginal background, immigrant status, visible minority, order, and panel), one factor at four levels (education), and one factor at five levels (region). The main explanatory variable is Panel.

4.1 Method I

In order to apply standard techniques in survival analysis, we divide the subjects by number of periods in unemployment. Hence, in the first analysis, we group the observations by number of unemployment periods and each duration corresponds to a single subject. The Table 4.2 shows the distribution of people having different number of unemployment periods in both Panels.

From Table 4.2 we can see that the number of observations goes down as the number of periods in unemployment increases, as has to be expected. We did not analyze persons with five and more unemployment periods since the number of observations is too small in such cases and reporting them would violate the confidentiality requirements of Statistics Canada.

Number of Periods	Panel 4 n	Panel 5 n
1	4 361	4 274
2	1 718	1 597
3	808	767
4	389	339
5 and more	268	231
Total	7 544	7 208

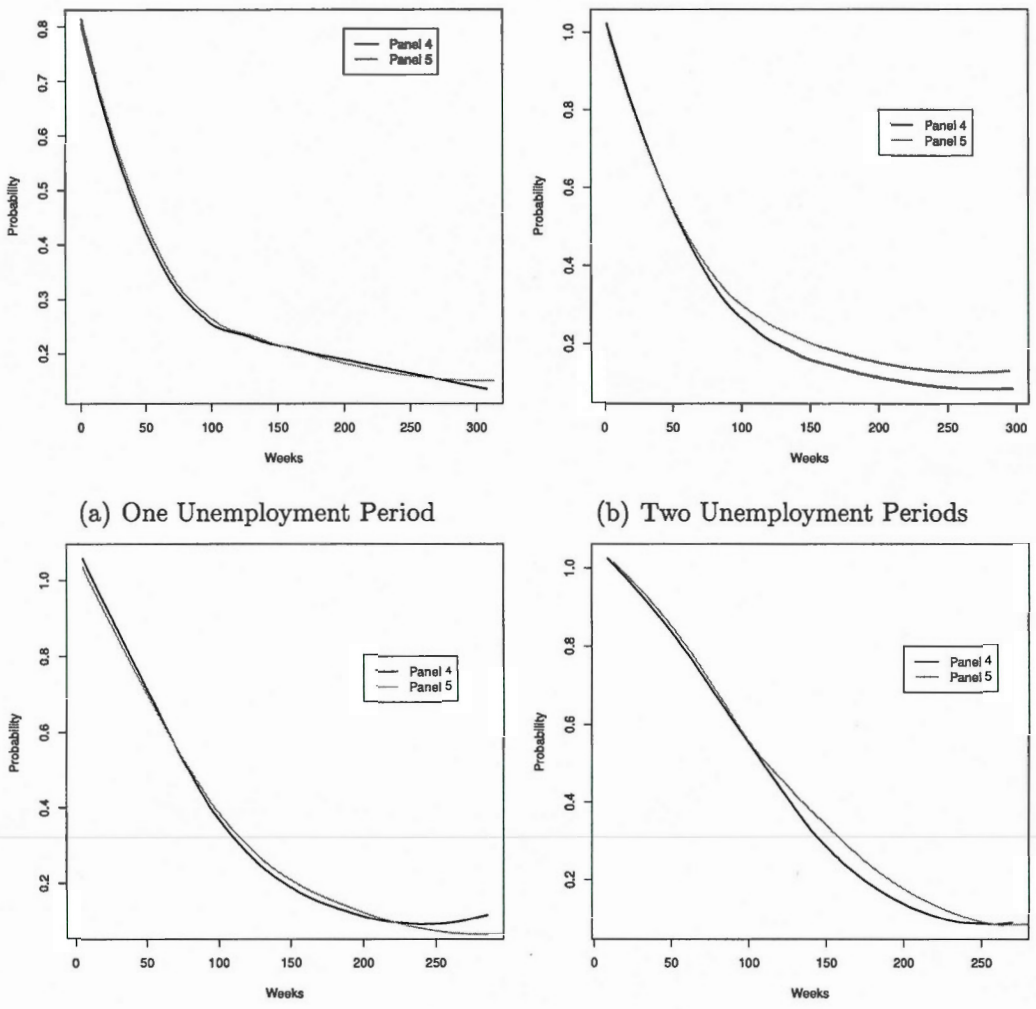
Table 4.2: The distribution by number of unemployment periods in both panels

The two distributions in Table 4.2 look identical but we performed also a formal chi-square test for independence. As the $p - value$ is 0.30 we cannot reject the null hypothesis that the distribution of periods in unemployment is the same in Panel 4 and Panel 5. This suggests that we could eventually consider to pursue in the direction pointed in Section 3.5, Remark (iii).

4.1.1 Non parametric estimation

In this section, the standard estimator of the survival function, proposed by Kaplan and Meier (1958) is used to estimate the survival and cumulative hazard function for our data (See Chapter III, Section 3.2).

We test for the difference in survival functions using two ways: an eyeball test based on the survival curves plots and a G-rho family of tests to check for the difference between two or more survival curves. Both methodologies are described in Chapter III, Section 3.3.



(a) One Unemployment Period

(b) Two Unemployment Periods

(c) Three Unemployment Periods

(d) Four Unemployment Periods

Figure 4.1: Smoothed Kaplan-Meier product limit estimates of unemployment durations in Panel 4 and Panel 5.

Statistics Canada confidentiality agreement did not allow us to exhibit the Kaplan-Meier plots under the premise that it is possible to identify individual participants in the SLID Survey. On the other hand, we were allowed to produce smoothed survival plots using smoothing techniques (namely the LOESS regression function in R, see Appendix A). The smoothed survival plots can be

found in Figure 4.1.

As mentioned above, initially we computed the Kaplan-Meier estimator to see if there is a difference in the probability of being unemployed given that the observation is in Panel 4 or Panel 5. The smoothed plots reproduce-roughly-the behavior of the true Kaplan-Meier curves. We observe that the smoothed Kaplan-Meier plots are superimposed and have crossing points, and therefore the Cox model is not valid when comparing the two panels (by subject). On the other hand, we note that for people having $k = 2$ unemployment periods, when the number of weeks in unemployment attains 50 (roughly a year) the duration in unemployment for people in Panel 5 is systematically higher than for those in Panel 4.

Number of Periods	$p - values$
1	0.164
2	0.677
3	0.836
4	0.371

Table 4.3: G-rho test (rho=1)

As mentioned in Chapter III, Section 3.3, the G-rho Statistics can test if there is a difference between two or more survival curves. The null hypothesis is that there is no difference between the survival curves versus the alternative that there exists a difference. Table 4.3 gives the $p - values$ for the G-rho test by number of periods in unemployment. As we can see, in all cases we do not reject the null hypothesis, that is, we find that there is no difference in unemployment durations between Panel 4 and Panel 5 (for people having the same number of

unemployment periods).

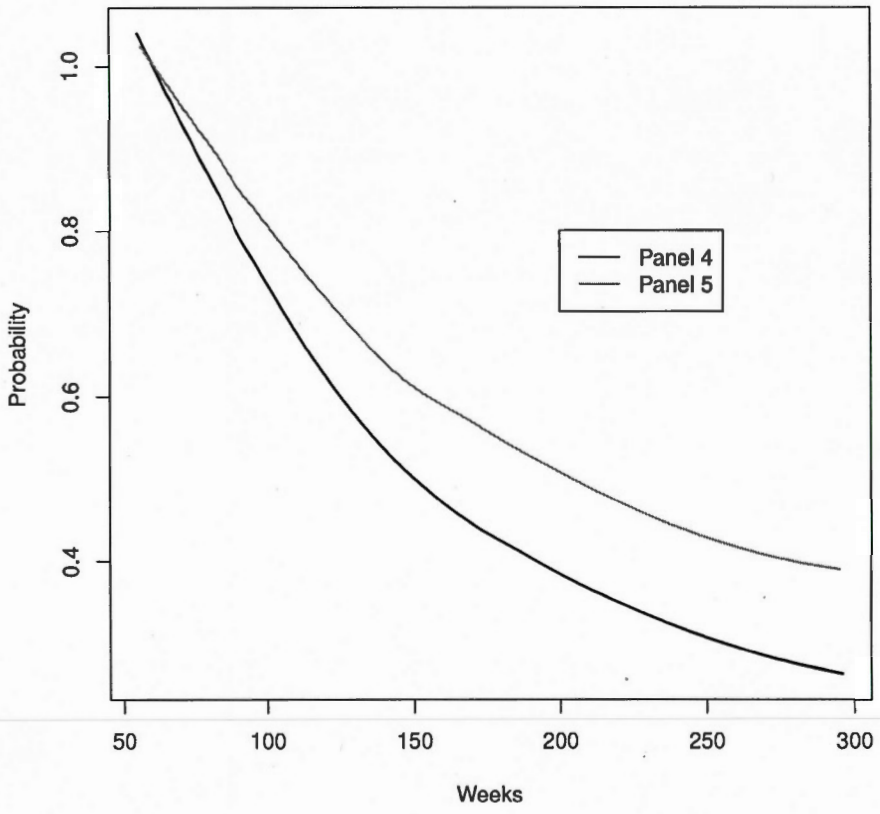


Figure 4.2: Conditional smoothed Kaplan-Meier product limit estimates of unemployment durations in Panel 4 and Panel 5 for people with $k = 2$ unemployment periods and more than 50 weeks in unemployment.

On the other hand, we noticed in Figure 4.1-b that no crossing point occurs on the Kaplan-Meier curves for Panel 4 and Panel 5 beyond 50 weeks in unemployment; hence, we decided to treat this case separately. Figure 4.2 shows the smoothed Kaplan-Meier curves for people having $k = 2$ unemployment periods and a total of more than 50 weeks in unemployment. For these conditional survival curves, the G-rho test gives a $p - value = 0.0052$, and we can reject the null hypothesis. Therefore, there exists some evidence that there is a significant difference between

Panel 4 and Panel 5 for this subgroup of people. We can also conclude from the same figure that the proportionality assumption seems to hold in this case. Therefore, we performed additional analyses on this group of people.

To conclude, as mentioned in the Introduction, we want to estimate the difference in unemployment duration per person for two groups of individuals, namely, individuals followed during the time window 2002-2007 (period with no crisis) versus individuals followed during the time window 2005-2010 (period covering the crisis). We also want to determine the set of covariates that could explain the difference in the unemployment duration between these two periods. In our analysis, the interpretation of the coefficients and the comparison of survival curves is based on formulas (3.14) and (3.15). In what follows, we start by trying to answer these questions using the dataset by subject (Method I, Chapter II).

4.1.2 Semi-parametric estimation

From Section 4.1.1 we can see that the Cox model is not applicable for all the dataset arranged by subject and conditioned by number of periods in unemployment. That is, from Figure 4.1 we observe that one can not use the Cox model since those graphics violate the proportionality assumption. However, we also observe from Figure 4.2 that conditional on $k = 2$ unemployment periods and more than 50 weeks in unemployment, the proportionality assumption holds and that the Cox model is applicable for this subgroup of persons. We refer to this Cox analysis as "Cox Model-I".

In Tables 4.4 and Table 4.5, the 596 subjects to which the Cox-Model-I applies are divided for each fixed and dynamic covariate. Notice that the distribution of persons is similar to what was presented in Table 2.8 and Table 2.9. For Panel 4

we have around 149 right censored unemployment durations out of 305 and for Panel 5 we have 174 out of 291 respectively. These proportions are quite large, on the other hand, they are partly due to the fact that we work conditionally, so the results should be understood accordingly. Note that it is not possible to provide detailed information per Panel since this would violate the confidentiality agreement for some categories.

Variable	Description	Database n
Sex	Male	267
	Female	329
	Total	596
Aboriginal background	Yes	30
	No	566
	Total	596
Immigrant	Yes	75
	No	521
	Total	596
Visible minority	Yes	57
	No	539
	Total	596

Table 4.4: Number of observations for the fixed factors. The first class in the list is the baseline category in our analysis.

Fitting the proportional hazards model given in Chapter III for the unemployment duration dataset described in Chapter II entails estimating the unknown coefficients of the explanatory variables such as age and the ones mentioned in

Tables 4.4 and 4.5, in the linear component of the model, β_l for $l = 1, ..., L$.

Variable	Description	Database n
Region	Atlantic	145
	Quebec	107
	Ontario	162
	Prairies	122
	British Columbia	60
	Total	596
Level of Education	No high school diploma	110
	High school diploma	168
	Non-university postsecondary certificate	205
	University degree or certificate	113
	Total	596

Table 4.5: Number of observations for dynamic factors. The first class in the list is the baseline category in our analysis.

The β-coefficients in the proportional hazards model, which are the unknown parameters in the model, have been estimated using the "Survival" package in R.

For convenience, we write in abbreviated form

$$\mathbf{x} = \begin{bmatrix} \text{Panel} \\ \text{Age} \\ \text{Sex} \\ \text{Aboriginal status} \\ \text{Immigrant} \\ \text{Visible minority} \\ \text{Region} \\ \text{Education Level} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ \mathbf{x}_7 \\ \mathbf{x}_8 \end{bmatrix} \tag{4.1}$$

where $\mathbf{x}_7$ has four components and $\mathbf{x}_8$ has 3 components (see Table 4.1).

As mentioned in Chapter III, the fitted proportional hazards model for the i-th individual is then

$$\hat{h}_i(t) = \exp\left(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i\right) \hat{h}_0(t) \tag{4.2}$$

where the subscript i on an explanatory vector denotes the value of that vector for the i-th individual, $i = 1, ..., n$ and $\hat{h}_0(t)$ is the estimated baseline hazard function.

To determine which of the eight explanatory variates are significant, a number of different models have been fitted, and the results compared. We use the ANOVA function in R for comparing the different models and final model selection (see Appendix A. Model selection, Table A.1.).

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Panel5	-0.388	0.679	0.127	< 0.001	0.533	0.863
Age	-0.033	0.967	0.004	< 0.001	0.960	0.974
n = 596						
Number of events = 273						

Table 4.6: Cox regression, "R" output for Method I when conditioning for $k = 2$ and more than 50 weeks in unemployment

Table 4.6 gives the estimated coefficients of the explanatory variables, the standard errors, $p - value$, and the limits for the confidence intervals for $\exp(\beta_i)$. Notice that the only significant covariates for these analyses are the Panel and the age of the person. In addition, the likelihood ratio test rejects the null hypothesis that all the β's are zero. The corresponding analysis of deviance is given in Appendix A.

The final model shown in Table 4.6 indicates significant effects at the 5% level for age (older persons have a smaller hazard, so tend to have jobless spells of longer duration) and being in Panel 5 (persons unemployed in the time span from 2005-2010 tend to have unemployment periods of longer duration than people belonging to the time span 2002-2007 for the same age). Remember that this conclusion is valid for people with $k = 2$ unemployment periods and more than 50 weeks in unemployment.

The exponential coefficients in the second column of Table 4.6 are interpretable as multiplicative effects on the hazard. Thus, for example, holding the other covariates constant (i.e. Panel), an additional year of age reduces the hazard of

unemployment duration by a factor of $\exp^{(-0.033)} = 0.967$ on average - that is, by 3.3 percent, and thus the survival probability of staying in unemployment is higher. For this reason, older persons tend to have jobless spells of longer duration.

For the same model, for people with the same age, the estimated hazard for people in Panel 5 is $\exp^{(-0.388)} = 0.679$ times that of the control Panel (Panel 4). So, the hazard decreases by 32.1 percent, and therefore, people belonging to Panel 5 tend to have unemployment periods of longer duration compared to people belonging to Panel 4.

4.1.3 Residuals

As mentioned in Chapter III, we use the Cox-Snell residuals to check the model adequacy. In what follows, we show the residuals for the Method I.

Recall that the Cox-Snell residuals, r_{C_i}, have properties that are quite dissimilar to those of residuals used in linear regression analysis, for example, as they should behave like a random sample from an exponential distribution of parameter 1. Indeed, Figure 4.3 seems to indicate that the true cumulative hazard function conditional on the covariates of each of the models has an exponential distribution. Remember that Figure 4.3 is related the model corresponding to Method I where we compared the survival in Panel 4 and Panel 5 for people with $k = 2$ unemployment periods and more than 50 weeks in unemployment.

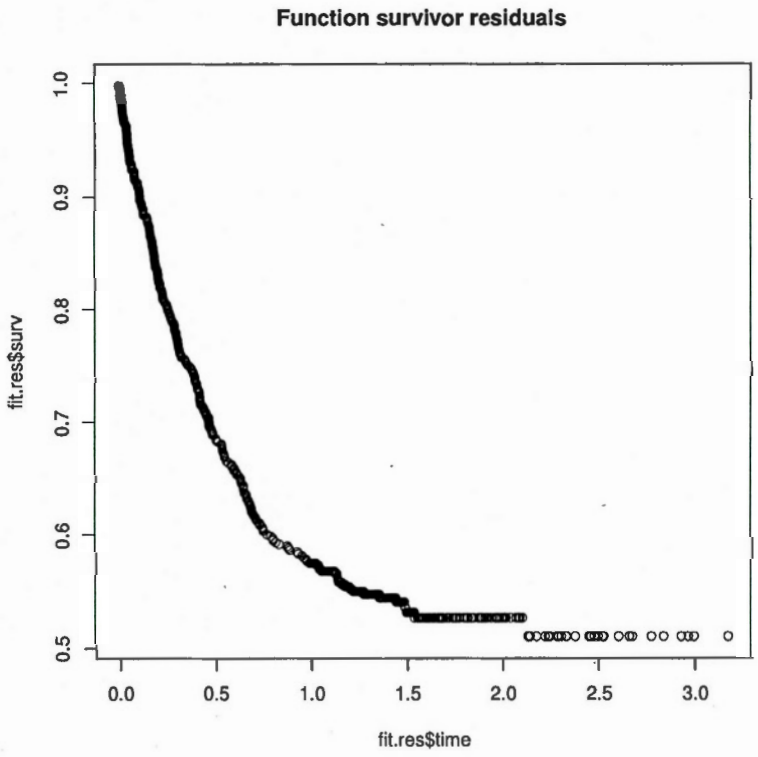


Figure 4.3: Method I. Residual Analysis.

4.2 Method II

This method consists in analyzing unemployment durations per observation instead of by subject (see Chapter II, section 2.4, Method II). Panel 4 includes 13 300 observations and Panel 5 includes 12 429 for a total of 25 729 observations. The complete observations distribution by category is shown in Table 2.11. In addition to the variables in Table 2.11, the factor Order is added for each Panel. Remember that Order=0 means that the observation corresponds to the first period in unemployment and Order=1 denotes any other observation. The distribution of the number of observations for this new factor are shown in Table 4.7.

	Panel 4 (2002-2007)	Panel 5 (2005-2010)	Total
Order = 0	7 544	7 208	14 752
Order = 1	5 756	5 221	10 977
Total	13 300	12 429	25 729

Table 4.7: The distribution of the factor *Order*.

Although the distribution seems quite similar in both panels, a formal chi-square test rejects the null hypothesis of homogeneity ($p-value$ is 0.043). So, it appears that, during the crisis, there was a higher proportion of first unemployment durations, duration that could correspond to people who had not experienced unemployment ever before.

The remainder of this section is organized as the section for Method I. Initially, we compared the Kaplan-Meier estimate and, in this case, the Cox model seemed applicable. So, we performed the semi-parametric estimation of the hazard function. We finish this section by assessing the goodness of fit of the proposed model.

4.2.1 Non parametric estimation

In this section, we fit a Kaplan-Meier estimator to 13 300 observations in Panel 4 and 12 429 observations in Panel 5. Figure 4.3 shows the smoothed Kaplan-Meier estimates.

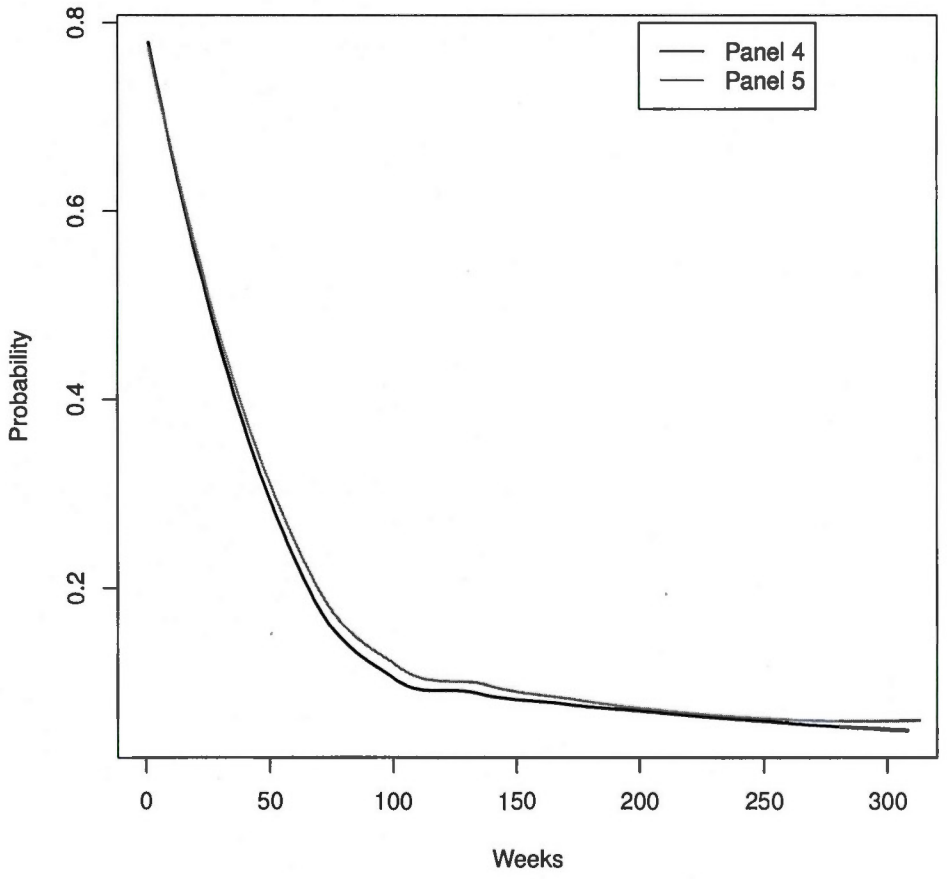


Figure 4.4: Smoothed Kaplan-Meier product limit estimates of unemployment durations of Panel 4 and Panel 5 by observation, Method II.

We remark that in this case, the curve of Panel 5 is always slightly above the curve corresponding to Panel 4. The $G - rho$ test rejects the null hypothesis of equality (the $p - value$ is equal to 0.03). We conclude that there is a difference in the probability $S(t)$ of being in unemployment longer than t when we compare observations from Panel 4 and Panel 5, and this probability is higher for Panel 5.

Notice that the analysis which is controlling for the number of periods in unemployment by grouping the data by subject generates different conclusions than the

one where the same subject can be considered more than once. In what follows, we give the results when fitting the Cox model for this non-grouped data.

4.2.2 Semi-parametric estimation

We consider all the covariates given in Table 4.1 for which we fit a Cox regression. The R ouput is reproduced in Table 4.8. Note that the variable age and the factors visible minority and order have highly statistically significant coefficients, while the coefficient for Panel 5 is not significant. However, when we consider the interaction between the factors Panel and Region the terms related to Panel 5 $\times$ Ontario and Panel 5 $\times$ BC are significant at 10 percent. The likelihood ratio test reveals that the null hypothesis that all of the β's are zero is rejected.

We can see immediately the difference between the results for Method I and Method II by comparing Table 4.6 and 4.8. Method II considers three more factors (region, visible minority and Order) and the interaction (Panel $\times$ region) to be important when we try to compare the unemployment duration for a period covering the financial crisis versus a period preceding the crisis. The analysis of deviance for the model presented in Tables 4.B.2 is in the Appendix B.

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Panel 5	0.022	1.022	0.02836	0.441	0.967	1.080
Age	-0.016	0.984	0.0004	0.000	0.983	0.985
QC	-0.006	0.994	0.0308	0.838	0.935	1.056
ON	-0.011	0.990	0.0273	0.688	0.938	1.043
AB, SK, MB	0.082	1.086	0.0279	0.003	1.028	1.147
BC	0.001	1.001	0.0387	0.975	0.928	1.080
Vis. Min.	0.111	1.118	0.0278	0.000	1.059	1.180
Order=1	0.210	1.234	0.0143	0.000	1.200	1.270
Panel5×QC	-0.012	0.988	0.0438	0.777	0.906	1.076
Panel5×ON	-0.108	0.897	0.0389	0.004	0.831	0.968
Panel5×AB,SK,MB	-0.046	0.955	0.0397	0.247	0.883	1.032
Panel5×BC	-0.093	0.911	0.0554	0.094	0.817	1.016
n = 25 729						
Number of events = 21 175						

Table 4.8: Method II. Cox regression.

Table 4.8 indicates significant effects for Age, that is, older persons have a smaller hazard, so tend to have longer periods in unemployment holding all other covariates constant. For visible minorities, we found that observations which are not part of a visible minority group have a higher hazard, and therefore, they tend to have shorter periods in unemployment than observations of visible minorities, when holding all other covariates constant. For the factor Order, our results show that observations for the second and subsequent unemployment periods have a greater hazard of leaving unemployment at any given time, and consequently, a second, third, etc., unemployment period tends to be shorter than a first unemployment period, which makes practical sense.

In the following analysis and computing we use the full decimal values and not the rounded ones presented in Table 4.8. The results suggest that persons from Ontario belonging to Panel 4 have a higher risk of staying in unemployment than persons from the other regions. In particular they have a 9% higher risk of staying in unemployment than persons from the Prairies and around 1% higher risk than persons from Quebec or British Columbia respectively. The effect stays qualitatively the same for the period covering the crisis, however, persons from Ontario have a 15% higher risk of staying in unemployment than persons from the Prairies, 10% higher risk than persons from Quebec, and 3% higher risk than those from British Columbia, respectively.

For the Prairies the results are opposite to the Ontario ones, that is, during the period preceding the crisis a person from the Prairies tended to have shorter unemployment periods than all other regions. In particular, they had 5% less risk of staying in unemployment than persons from Quebec, 16% less risk than persons from Ontario, and 13% less than British Columbia respectively. During the crisis, persons from the Prairies have 9% less risk of staying in unemployment than persons from Quebec, 9% less risk than persons from Ontario, and 8% less risk than persons from British Columbia respectively. Hence, during the crisis the region of the Prairies performed better than all others.

To summarize, Method II includes more factors as significant explanatory variables than Method I. Our analysis suggests that observations belonging to members of a minority group have been more affected by the crisis than those corresponding to members of a no visible minority group. This result suggests a possible discrimination in the labour market, that is, in difficult periods, those who are not members of visible minorities tend to find new employment faster

than members of visible minorities. Of course, this could be due to other related factors like years of service in the Canadian labour market, e.g. Finally, observations from Ontario seem to be longer (most affected by the crisis) and observations from the Prairies region to be shorter (less affected).

4.2.3 Residuals

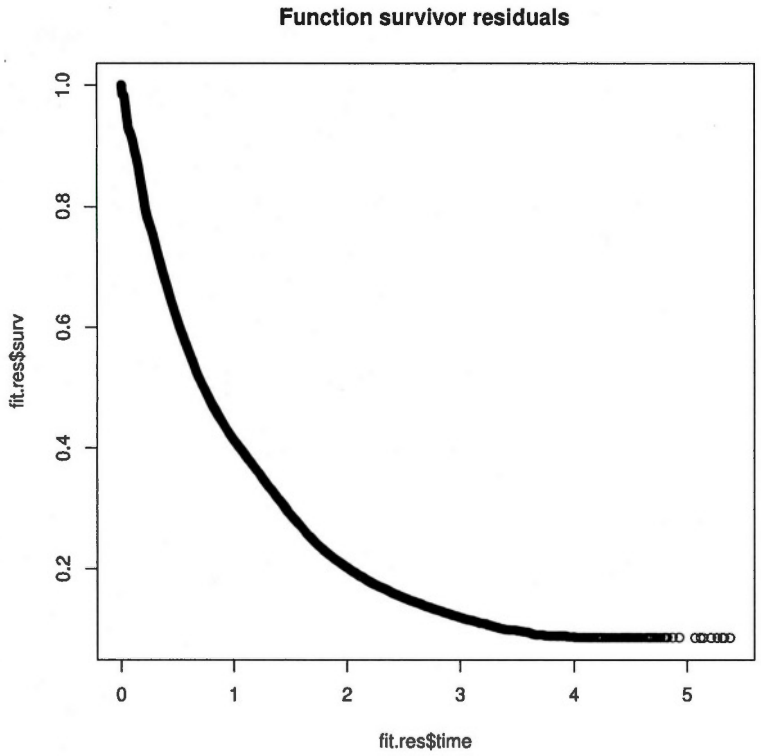


Figure 4.5: Method II. Residual Analysis

In Figure 4.5 we give the Cox-Snell residuals for Method II. In the context of this model with 5 explanatory covariates and one interaction, the Cox-Snell residuals seem to follow an exponential distribution, and therefore we conclude that our model captures the essential covariates explaining the unemployment duration for the model analyzing observations instead of subjects.

4.3 Other analyses

From Figure 4.1 we conclude that the Cox model is not applicable when comparing the behaviour of unemployment duration between the two panels for the same number of unemployment periods, except in the conditional case commented above. However, we can determine which covariates explain the unemployment duration at the interior of each of the panels for a given number of unemployment periods.

After comparing the two methodologies mentioned in Chapter II we focus on finding the set of covariates that have affected the unemployment duration for the time windows covered by each panel. In this case, we analyze each panel conditioning on the number of unemployment periods ($k = 1, 2, 3$) per subject ($i = 1, .., n$) in each panel (Panel 4 and Panel 5.). We summarize our results in the tables that follow Section 4.3.1.

4.3.1 Semi-parametric estimation

We analyze the data by the number of unemployment periods. **For $k = 1$ unemployment periods**, Table 4.9 gives the final model for Panel 4 and Table 4.10 gives the final model for Panel 5. The analysis of deviance for these models is given in the Appendix B, Table B.3 and Table B.4, respectively.

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Age	-0.024	0.976	0.001	< 0.001	0.974	0.978
Sex	-0.173	0.841	0.036	< 0.001	0.783	0.903
n = 4 361						
Number of events = 3 048						

Table 4.9: Cox regression for Panel 4 and $k = 1$ unemployment periods.

Table 4.9 shows that for Panel 4 the variable age and the factor sex are significant. Table 4.10 shows that for Panel 5 the variable age, the factors sex, region and education are significant. In both cases, the likelihood ratio test rejects the null hypothesis that all of the β's are zero.

We remark that the effect of age on the unemployment duration is quite similar for both panels (seems equal because of the rounding process). As before, holding all factors constant, older persons have a smaller hazard, so tend to have jobless spells of longer duration; unemployed women tend to have unemployment periods of longer duration in the period before and during the crisis, when holding all other covariates constant. Women belonging to Panel 5 have a 19% higher risk of staying in unemployment than men, while in the case of Panel 4 this risk is higher by 15%. Therefore, it seems that, for women, the risk of staying in unemployment increased by 4% during the crisis.

Now, we focus only in Table 4.10, that is, for persons belonging to Panel 5. In addition to the variable age and the factor sex, the factors region and education are significant. Note that holding all other covariates constant, persons from Ontario have 19% more risk of staying in unemployment than persons from the

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Age	-0.024	0.976	0.001	< 0.001	0.974	0.978
Women	-0.213	0.808	0.037	< 0.001	0.752	0.868
QC	-0.098	0.906	0.060	0.105	0.805	1.021
ON	-0.148	0.862	0.054	< 0.005	0.776	0.958
AB, SK, MB	0.066	1.068	0.054	0.224	0.960	1.188
BC	-0.076	0.927	0.073	0.294	0.803	1.068
Educ2	0.239	1.270	0.056	< 0.001	1.138	1.418
Educ3	0.489	1.631	0.056	< 0.001	1.460	1.822
Educ4	0.413	1.511	0.064	< 0.001	1.333	1.713
n = 4 274						
Number of events = 3 037						

Table 4.10: Cox regression for Panel 5 and $k = 1$ unemployment periods.

Atlantic region.

Results from Table 4.10 show that all higher education levels are highly significant at the 1% level, implying that persons with higher education levels tend to have a smaller unemployment period than persons no high school diploma. In particular, persons with high school diploma, persons with a non-university certificates, and persons with university degrees have respectively, 27%, 63%, and 51% less risk of staying in unemployment than persons from the baseline category (no high school diploma). Hence, more skilled workers tend to find a job faster during the crises than low skilled workers. However, persons with no university certificate have 12% less risk than persons with a university degree.

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Age	-0.016	0.984	0.002	< 0.001	0.980	0.988
Sex	-0.142	0.867	0.057	0.013	0.775	0.970
QC	0.105	1.110	0.093	0.260	0.926	1.332
ON	0.052	1.054	0.083	0.527	0.896	1.240
AB, SK, MB	0.253	1.287	0.084	0.003	1.092	1.518
BC	0.136	1.146	0.119	0.226	0.919	1.430
Educ2	0.082	1.086	0.088	0.353	0.913	1.291
Educ3	0.288	1.333	0.089	0.001	1.120	1.587
Educ4	0.114	1.120	0.108	0.290	0.907	1.385
n = 1718						
Number of events = 1238						

Table 4.11: Cox regression, Panel 4 and $k = 2$ unemployment periods.

For $k = 2$ unemployment periods, Table 4.11 gives the final model for Panel 4 and Table 4.12 gives the final model for Panel 5. The analysis of deviance for these models is presented in the Appendix B in Table B.5 and Table B.6, respectively. In this case we observe that for Panel 4 the variable age and the factors sex, region, and education are significant. For Panel 5, only the variable age and the factor education are significant. The likelihood ratio test rejects the null hypothesis that all the β's are zero.

For the models presented in Table 4.11 and Table 4.12 the variable age has the same qualitative effect as in the previous models. For Panel 4, holding all other factors constant, women tend to have jobless spells of longer duration than men. Namely, women have 13.3% more risk of staying in unemployment than men. Persons from the Prairies region tend to have jobless spells of smaller dura-

tion than persons from the Atlantic region, explicitly, persons from the Prairies have 29% less risk to stay in unemployment than persons from the Atlantic region.

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Age	-0.013	0.987	0.020	< 0.001	0.9834	0.9911
Educ2	0.239	1.270	0.095	0.012	1.0535	1.5308
Educ3	0.295	1.342	0.095	0.002	1.1133	1.6189
Educ4	0.371	1.450	0.107	0.000	1.1759	1.7871
n = 1 597						
Number of events = 1 106						

Table 4.12: Cox regression, Panel 5 and $k = 2$ unemployment periods.

Again, education seems to play an important role during the crisis. In Panel 4, holding all other covariates constant, persons with a high school diploma or with a non-university certificate have (8% and 33% respectively) less risk to stay in unemployment than persons with no high school diploma. During the crisis (Panel 5), persons with a high school diploma, or a non-university diploma, or a university diploma have less risk to stay in unemployment than persons with no high school diploma (27%, 34%, and 45% less risk respectively).

Notice that, when we control for the number of unemployment periods, in this case when we take $k = 2$, persons with a university certificate are less at risk to stay in unemployment than persons with no university certificate (11%) or persons with a high school diploma (18%). Controlling for the number of unemployment periods k and analyzing by subject yields to the conclusion that higher education levels for persons with two unemployment periods tend to decrease the risk to stay in unemployment. Even if in Method II education was

a significant factor, the conclusion for the model from this section and from Method II are different. In Method II we conclude that having a university certificate increases the risk to stay in unemployment in comparison with the case of a non university certificate or high school diploma, which is somewhat counterintuitive from the economic point of view.

Finally, **for $k = 3$ unemployment periods** Table 4.13 gives the selected model for Panel 4 and Table 4.14 gives the model for Panel 5. As before, the deviance analysis tables are given in the Appendix B, namely in Table B.7 and Table B.8. For Panel 4 the variable age, the factors sex and visible minority are significant while for Panel 5 the only significant factor is the region. We were surprised that the age is no longer a significant covariate for people with three unemployment periods and belonging to Panel 5.

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
Age	-0.006	0.994	0.003	0.046	0.988	0.999
Sex	-0.230	0.794	0.086	0.007	0.670	0.940
Vis. Min.	0.394	1.483	0.191	0.039	1.019	2.159
n = 808						
Number of events = 547						

Table 4.13: Cox regression, Panel 4 and $k = 3$ unemployment periods.

For Panel 4, the variable age has the same qualitative effect as in the previous models. Women have approximately 20% more risk to stay in unemployment than men having the same age and belonging to the same visible minority group. Members of no visible minority group have 48% less risk of staying in

unemployment than members of a visible minority group. Hence, during the period preceding the financial crisis, women and members of visible minority groups tended to have longer unemployment periods.

For Panel 5, persons from the Prairies region have 27% less risk to stay in unemployment than persons from the Atlantic region. This is a result found in almost all models where the region factor is significant during the financial crisis. Since the coefficients for the other regions are not significant, we do not report the observed difference in these cases.

	coef	exp(coef)	s.e.(coef)	p-value	lower .95	upper .95
QC	0.167	1.18 2	0.132	0.205	0.9127	1.530
ON	-0.081	0.922	0.122	0.504	0.7253	1.171
AB, SK, MB	0.242	1.274	0.123	0.050	1.0002	1.622
BC	-0.207	0.813	0.186	0.264	0.5647	1.170
n = 767						
Number of events = 535						

Table 4.14: Cox regression, Panel 5 and $k = 3$ unemployment periods.

Finally, because the number of observations is lower for both Panels when $k = 4$, no covariate is significant in this case. We summarize all the previous ($k = 1, 2, 3$) results in Table 4.15.

k	Panel 4 (2002-2007)	Panel 5 (2005-2010)
1	Age and sex	Age, sex, region, and education level
2	Age, sex, region, and education level	Age and education level
3	Age, sex, and visible minority	Region

Table 4.15: Significant covariates for Panel 4 and Panel 5

4.3.2 Residuals and conclusion

Figure 4.6 refers to the residuals for each of the models computed for Panel 4 and Panel 5 controlling for the number k of unemployment periods. Since in all cases the residuals seem to behave like a random sample from an exponential distribution of parameter one, we conclude that our models capture the main covariates which explain unemployment duration for the periods before and during the financial crisis.

We summarize our findings of Section 4.3 in the following remarks.

i) Except the model for Panel 5 and $k = 3$ unemployment periods, the variable Age seems to be significant in all other models with the same qualitative effect: a one year increase yields a higher probability of staying in unemployment.

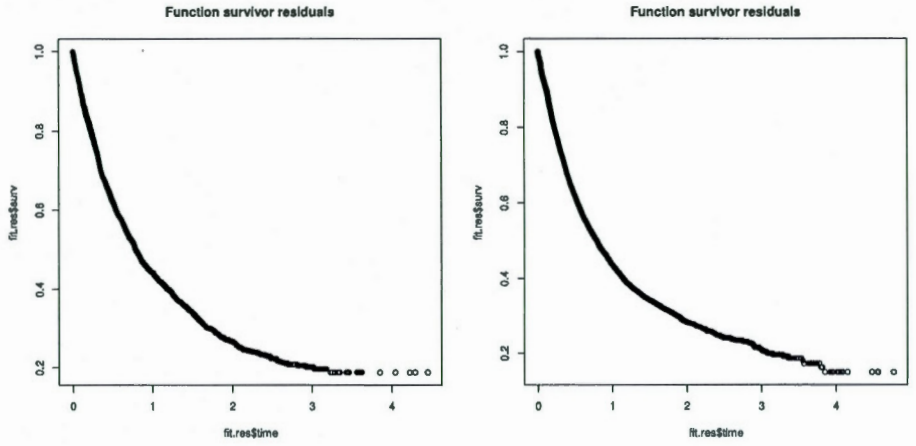
ii) The variable *Age* can not help us to explain the difference in unemployment duration between Panel 4 and Panel 5. Older people tend to have longer unemployment spells in periods with or without a crisis.

iii) In the period before the financial crisis, that is, for individuals belonging to Panel 4, to be a woman seems to be linked to a higher probability of staying in unemployment. Sex difference does not seem to have played an important role during the financial crisis for persons with $k = 2, 3$ unemployment periods.

iv) For Panel 4, members of a visible minority tend to have longer unemployment durations than persons who are not. This conclusion applies only for people having $k = 3$ unemployment periods.

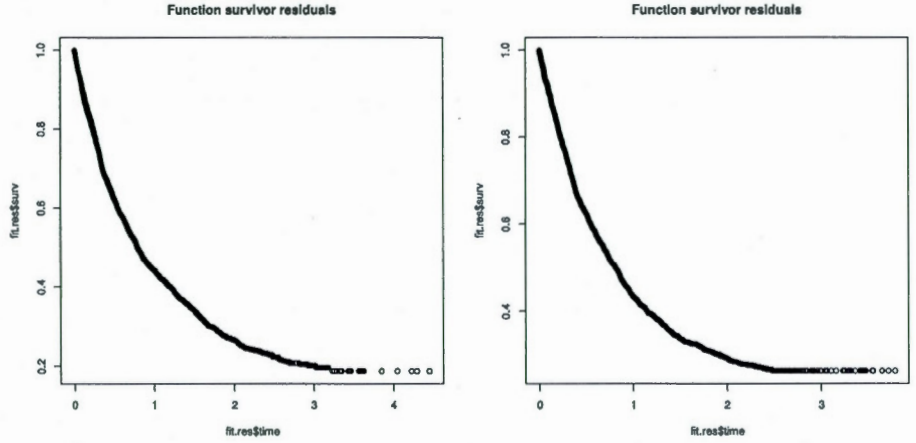
v) Higher education levels for individuals belonging to Panel 5 tend to decrease the duration in unemployment. So, higher education seems to have played a protective role during the financial crisis.

vi) The provinces of Alberta, Saskatchewan, and Manitoba (the Prairies region) perform better than all other regions in both periods, namely before and during the crisis. The province of Ontario was the most touched by the crisis, that is, persons from Ontario tend to haver higher risk of staying in unemployment during the crisis than all other regions.



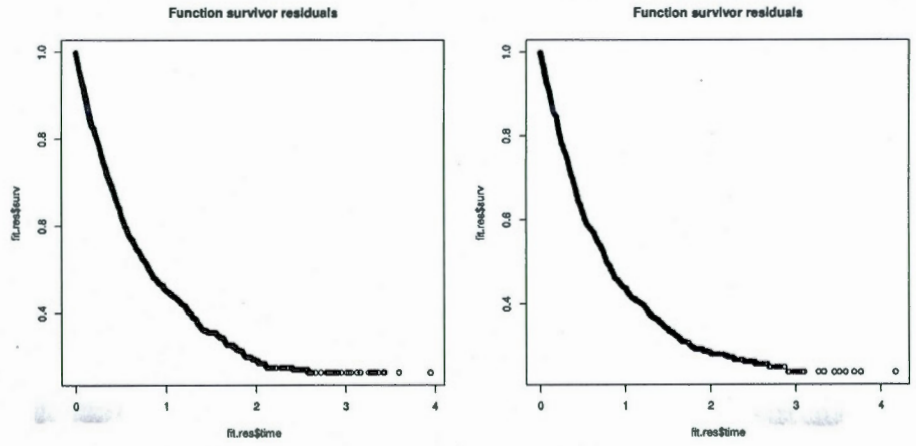
(a) Panel 4 and $k = 1$

(b) Panel 5 and $k = 1$



(c) Panel 4 and $k = 2$

(d) Panel 5 and $k = 2$



(e) Panel 4 and $k = 3$

(f) Panel 5 and $k = 3$

Figure 4.6: Residual Analysis for Panel 4 and Panel 5 for different $k = 1, 2, 3$ unemployment periods.

[Cette page a été laissée intentionnellement blanche]

CONCLUSION

In this Master Thesis we used the unweighted data from the Survey of Labour and Income Dynamics (SLID) to assess some factors affecting the duration of unemployment during the 2008 financial crisis in Canada. To analyse the 2008 financial crisis effect on the Canadian labour market we compared two panels from the SLID: a) Panel 4 which covers the time window from January 1st, 2002 to December 31st, 2007, and b) Panel 5 which covers the time window from January 1st, 2005 to December 31st, 2010. Moreover, we work conditionally, in the sense that we eliminate all observations with starting date before the beginning date of each of the panels, so the results should be understood accordingly. In the Method I analyses we also conditioned by the number of periods in unemployment.

We carried out our analysis using two different methods: I) an analysis based on the total unemployment duration for each individual in the study and II) an analysis based on duration times which ignores that the same person could correspond to more than once unemployment duration. The results differ between the two methods. We used the classical biostatistics statistical methodology of survival analysis in each case.

In Method I we conditioned by the number of unemployment periods. The non-parametric estimation for Method I does not allow us to clearly determine whether the unemployment duration is higher in Panel 4 than in Panel 5 (or vice-versa), while Method II reveals a slightly higher probability to stay in

unemployment for observations belonging to Panel 5 than for observations belonging to Panel 4. However, in Method I we can see a higher probability to stay in unemployment for individuals belonging to Panel 5, when we condition on two unemployment periods and more than 50 weeks in unemployment.

Further, in trying to assess which factors can affect unemployment duration it turns out that the significant explanatory variables (like personal characteristics) are different for both methods. Conditioning by observing two unemployment periods and more than 50 weeks in unemployment, the only significant variables in Method I are the variable age and the factor Panel. For Method II, in addition to the variable age and factor Panel, three additional factors are significant (Region, Visible Minority, and Order).

Additional to Method I and Method II analyses for comparing the two panels we also performed an analysis for each of the panels separately in order to assess which covariates determined the unemployment durations in their respective time windows. This analysis was done conditioning by number of unemployment periods $k = 1, 2, 3$. The sets of characteristics with strong effects on unemployment duration are remarkably similar in both cohorts. The most important findings are: i) higher education levels tended to decrease unemployment duration during the crisis period; ii) the Canadian regions were touched differently by the crisis, in particular, the provinces of Alberta, Saskatchewan and Manitoba seemed to perform better during the crisis than the other regions, while the province of Ontario seemed to be the most affected by the crisis; and iii) personal characteristics seemed to be important during the period with no crisis, that is, women and members of visible minorities tended to have longer unemployment periods.

Method II gives results which are consistent with the unemployment figures released by Statistics Canada and obtained through cross-sectional studies. This is not that surprising given that in Method II one deals with separate unemployment periods which ignore each subject's progression, they are "cross-sectional" in nature.

We have not discussed deeply the issue whether Method I is more appropriate than Method II statistically. Economically speaking, it seems more interesting to perform an analysis for the total duration in unemployment by subject than by observation. Statistically it seems that taking into account the possible dependence between observations should be important as well. Actually, resorting to recurrent event analysis techniques could be the most adequate tool in this context. This is beyond the scope of the present Master Thesis, but research into this area is, of course, of interest, especially in connection with longitudinal surveys such as SLID, where individuals can experience successive spells of unemployment.

Finally, there are no major discrepancies between our results (unweighted data) and the tendencies found by using the weighted data (see, e.g., Indicators of Well-Being in Canada, Unemployment Duration, Statistics Canada, 2013)[1]. Further research could address the issue of comparing the results obtained by these different methodologies.

This research was supported by funds to the Canadian Research Data Centre

[1]http://www4.hrsdc.gc.ca/.3ndic.1t.4r@-eng.jsp?iid=15

Network (CRDCN) from the Social Science and Humanities research Council (SSHRC), the Canadian Institute for Health Research (CIHR), the Canadian Foundation for Innovation (CFI) and Statistics Canada.

Although the research and analysis are based on data from Statistics Canada, the opinions expressed do not represent the views of Statistics Canada or the Canadian Research Data Centre Network (CRDCN).

APPENDIX A. LOESS REGRESSION

According to Statistics Canada, producing a step function could create confidentiality problems. For this reason, we needed to implement some additional methodology in order to be authorized to present some Kaplan-Meier plots. This methodology consists in applying LOESS regression techniques as described below.

LOESS regression is the method we used in order to smooth the original Kaplan-Meier curves. Like linear regression, the smooth curve is drawn in such a way as to have a minimal variance of the residuals or prediction error. The acronym LOESS is meant to represent the notion of **loc**al regr**ess**ion that provides a generally smooth curve, the value of which in a particular location along the x-axis is determined only by the points in the vicinity of that x point. The method makes no assumptions about the form of the relationship, and allows the form to be discovered using the data itself.

Let (x_i, y_i), $i = 1, ..., M$ be the data points. In LOESS one applies least squares regression locally. This leads to estimate the vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ which minimize

$$\sum_{i=1}^{n} W_{ki}(x) \left(y_i - \sum_{j=0}^{p} \beta_j x^j \right)^2,$$

where $W_{ki}(x)$ denote some local weights i.e. $W_{ki}(x) = 0$ for data points (x_i, y_i) such that x_i is far from x (e.g. $|x_i - x| > d_i$, d_i a window width, $i = 1, ..., n$).

In our case, we use the LOESS function implemented in R as follows:

i) we compute the Kaplan-Meier estimate;

ii) let x_i denote the time $(i = 1, ..., n)$;

iii) let y_i denote the value from the survivor function from the Kaplan-Meier estimate $(i = 1, ..., n)$;

iv) run the LOESS regression. We run an equi-weights regression with y_i as a dependent variable and x_i as the explanatory one $(i = 1, ..., n)$;

v) plot the predictions given by the function *predict* in R against x_i to build a smooth Kaplan-Meier function.

APPENDIX B. MODEL SELECTION

Some packages offer an automatic choice to select the important explanatory variables. The R package offers the function *step*, however, instead of using the automatic variable selection procedures, we followed the following general strategy.

i) The first step is to fit models that contain each of the variables one at a time. Then, compute the ANOVA (Table of deviance analysis) and compare the log likelihood of the last model versus the one of the null model to determine which variables are significant on their own.

ii) Fit a model with the all variables selected in i). In the presence of certain variables, others may cease to be important. We compute the deviance analysis table. Different models are fitted in this step, since changing the order of the covariates and for each we perform the deviance analysis. The deviance analysis reveals that some covariates are no longer significant and they are omitted from the following step.

iii) Variables that were not important on their own, and so were not under consideration in step ii), may become important in the presence of other variables. These variables are therefore added to the model from step i), one at a time. Then, one checks the deviance analysis table and verifies if this added variable is significant or not. This process may result in terms in the model determined at step ii) ceasing to be significant.

iv) From the selected variables in step ii) and iii), we compute models with interactions on the significant covariates, one at a time, look at the ANOVA

output and check if the interaction is or not significant. We proceeded for the interaction as in ii) and iii).

v) A final check is made to ensure that no term in the model can be omitted.

In what follows, we show the corresponding final deviance analysis output corresponding to the models fitted in Chapter IV.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-1580.0			
Panel	-1576.0	7.948	1	< 0.004
Age	-1533.1	85.784	1	< 0.001

Table B.1: Analysis of Deviance. Method I. Model presented in Table 4.6.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-197342			
Age	-196584	1516.715	1	< 0.001
Panel	-196581	5.523	1	0.019
Region	-196555	51.776	4	< 0.001
Vis. Min.	-196545	21.405	1	< 0.001
Order=1	-196438	212.937	1	< 0.001
Panel:Region	-196433	10.015	4	0.040

Table B.2: Analysis of Deviance. Method II. Model presented in Table 4.8.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-23372			
Age	-23132	480.929	1	< 0.004
Sex	-23120	22.605	1	< 0.004

Table B.3: Analysis of Deviance for Panel 4 and $k = 1$ unemployment periods. Model presented in Table 4.9.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-23268			
Age	-23043	448.486	1	< 0.001
Sex	-23032	23.519	1	< 0.001
Region	-23023	17.663	4	0.001
Educ	-22979	87.441	3	< 0.001

Table B.4: Analysis of Deviance. Panel 5 and $k = 1$ unemployment periods. Model presented in Table 4.10.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-8226.0			
Age	-8189.4	73.1374	1	< 0.001
Sex	-8186.2	6.3231	1	0.011
Region	-8180.6	11.2617	4	0.024
Educ	-8173.6	13.9131	3	0.003

Table B.5: Analysis of Deviance. Panel 4 and $k = 2$ unemployment periods. Model presented in Table 4.11.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-7335.0			
Age	-7312.3	45.411	1	< 0.001
Educ	-7305.2	14.187	3	0.002

Table B.6: Analysis of Deviance, Panel 5 and $k = 2$ unemployment periods. Model presented in Table 4.12.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-3216.7			
Age	-3215.3	2.8871	1	0.089
Sex	-3211.6	7.3201	1	0.007
Vis. Min.	-3209.2	4.7773	1	0.029

Table B.7: Analysis of Deviance, Panel 4 and $k = 3$ unemployment periods. Model presented in Table 4.13.

	loglik	Chisq	Df	Pr(> |Chi|)
NULL	-3118.4			
Region	-3112.9	11.123	4	0.02522 *

Table B.8: Analysis of Deviance, Panel 5 and $k = 3$ unemployment periods. Model presented in Table 4.14.

BIBLIOGRAPHY

Aidt, T. S. and Tzannatos, Z. (2008). Trade unions, collective bargaining and macroeconomic performance: a review. *Industrial Relations Journal*, *39*(4), 258–295.

Amela, K., Nachum, G. and Niels, V. (2012). *Measuring Labour Markets in Canada and the United States.* Fraser Institute.

Bergevin, P. (2008). Canada and the United States: The global financial crisis and its impact on Canada. *Parliament of Canada.*

Boudreau, C. and Lawless, J. F. (2006). Survival Analysis Based on the Proportional Hazards Model and Survey Data. *The Canadian Journal of Statistics*, *34*(2), 203–216.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* Chapam.

Corak, M. and Heisz, A. (1995). The duration of unemployment: a user guide. *Statistics Canada Working Paper*, *84*.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Hajducek, D. M. and Lawless, J. F. (2012). Duration analysis in longitudinal studies with intermittent observation times and losses to followup. *The Canadian Journal of Statistics*, *40*(1), 1–21.

Harrington, D. P. and Fleming, T. R. (1982). A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*, *69*(3), 553–566.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis. Techniques for Censored and Truncated Data.* Springer.

Lancaster, T. and Nickell, S. (1980). The Analyses of Re-employment Probabilities. *Journal of the Royal Statistical Society, Series A (General)*, 141–165.

Megginson, W. L. and Netter, J. M. (2001). From state to market: A survey of empirical studies on privatization. *Journal of economic literature*, *39*(2), 321–389.

Nickell, S. (1979). Estimating the probability of leaving unemployment. *Econometrica: Journal of the Econometric Society*, 1249–1266.

OECD. (2013). *Employment Outlook, 2013*. OECD.

Palda, F. (2000). Some deadweight losses from the minimum wage: the cases of full and partial compliance. *Labour Economics*, *7*(6), 751–783.

Rollin, A.-M. (2012). *Measures of Employment Turnover Post 2000: Gross Employment Gains and Losses Versus Net Employment Change*. Statistics Canada.

Stigler, G. J. (1946). The economics of minimum wage legislation. *The American Economic Review*, *36*(3), 358–365.

Zorn, L., Wilkins, C. and Engert, W. (2009). Bank of Canada liquidity actions in response to the financial market turmoil. *Bank of Canada Review*, 7–26.